SITA H. VERMEULEN

Genetic epidemiology of homocysteine and related diseases

SITA VERMEULEN

Genetic epidemiology of homocysteine and related diseases

Vermeulen, (Sita) Hendrika
Genetic epidemiology of homocysteine and related diseases
Thesis Radboud University Nijmegen with summary in Dutch

ISBN 978-90-9024401-3

Coverdesign and book lay-out: MWOntwerp, www.mwontwerp.nl

Printed by Ponssen & Looijen BV, Ede

Genetic epidemiology of homocysteine and related diseases

Een wetenschappelijke proeve op het gebied van de Medische Wetenschappen

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op woensdag 30 september 2009
om 13.30 uur precies

door

HENDRIKA HENRIETTE MARIA VERMEULEN

geboren op 15 juni 1978
te Appeltern

Promotores:	Prof. dr. L.A.L.M. Kiemeney Prof. dr. A.R.M.M. Hermus
Copromotores:	Dr. M. den Heijer Dr. H.J. Blom
Manuscriptcommissie:	Prof. dr. N.V.A.M. Knoers (voorzitter) Prof. dr. H. Snieder (UMCG, Groningen) Prof. dr. F.R. Rosendaal (LUMC, Leiden)

Financial support by the Netherlands Heart Foundation for the publication of this thesis is gratefully acknowledged. In addition, Orphan Europe is gratefully acknowledged for financial support for the printing of this thesis.

Table of contents

Part 2: Genetic epidemiological designs and analyses

Abbreviations

AdoHcy	S-adenosylhomocysteine
AdoMet	S-adenosylmethionine
AHCY	S-adenosylhomocysteine hydrolase
ALDH1L1	aldehyde dehydrogenase 1 family, member L1
AMD1	adenosylmethionine decarboxylase 1
AMT	aminomethyltransferase
APEX	arrayed primer extension
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
B12	vitamin B_{12} (cobalamin)
B2	vitamin B_2 (riboflavin)
B6	vitamin B_6 (pyridoxine)
BHMT	betaine-homocysteine S-methyltransferase
BHMT2	betaine-homocysteine methyltransferase 2
bp	basepair
CBS	cystathionine-beta-synthase
CH_2THF	5,10-methylenetetrahydrofolate
CH_3THF	5-methyltetrahydrofolate
CNV	copy-number variation
COMT	catechol-O-methyltransferase
COQ3	coenzyme Q3 homolog, methyltransferase
CTH	cystathionase
CUBN	cubilin (intrinsic factor-cobalamin receptor)
CVD	cardiovascular disease
Cys	cysteine
Cysta	cystathionine
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
DNMT1	DNA (cytosine-5-)-methyltransferase 1
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta
FCH	familial combined hyperlipidemia
FOLH1	folate hydrolase (prostate-specific membrane antigen)1
FOLR1	folate receptor 1 (adult)
FOLR2	folate receptor 2 (fetal)
FOLR3	folate receptor 3

FPGS	folylpolyglutamate synthase
FTCD	formiminotransferase cyclodeaminase
FUT2	fucosyltransferase 2
GAMT	guanidinoacetate N-methyltransferase
GART	phosphoribosylglycinamide formyltransferase
GGH	gamma-glutamyl hydrolase
GNMT	glycine N-methyltransferase
GWA	genome-wide association
Hcy	homocysteine
Hhcy	hyperhomocysteinemia
HPLC	high-pressure liquid chromatography
HWE	Hardy-Weinberg equilibrium
ICMT	isoprenylcysteine carboxyl methyltransferase
IMP	inosine monophosphate
LD	linkage disequilibrium
LOD	logarithm of odds
MAF	minor allele frequency
MAT	methionine adenosyltransferase
MAT1A	methionine adenosyltransferase I, alpha
MAT2A	methionine adenosyltransferase II, alpha
MDR	multifactor dimensionality reduction
Met	methionine
METTL1	methyltransferase-like gene 1
MGMT	O-6-methylguanine-DNA methyltransferase
MTHFD	methylenetetrahydrofolate dehydrogenase
MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2
MTHFR	methylenetetrahydrofolate reductase
MTHFS	5,10-methenyltetrahydrofolate synthetase
MTR	methionine synthase
MTRR	methionine synthase reductase
MTs	AdoMet-dependent methyltransferases
NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)
NNMT	nicotinamide N-methyltransferase
NOS1	nitric oxide synthase 1
NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)
NOS3	nitric oxide synthase 3 (endothelial cell)
NTD	neural tube defect
OR	odds ratio

PCMT1	protein-L-isoaspartate (D-aspartate) O-methyltransferase
PCR	polymerase chain reaction
PE	pulmonary embolism
PON1	paraoxonase 1
PPARG	peroxisome proliferator-activated receptor gamma gene
PRMT1	protein arginine methyltransferase 1
PRMT2	protein arginine methyltransferase 2
QTL	quantitative trait locus
RBC	red blood cell
RFC	reduced folate carrier
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RNMT	RNA (guanine-7-) methyltransferase
ROS	reactive oxygen species
RVT	recurrent venous thrombosis
SAHH	S-adenosylhomocysteine hydrolase
SARDH	sarcosine dehydrogenase
SHMT	serine hydroxymethyltransferase
SHMT1	serine hydroxymethyltransferase 1
SHMT2	serine hydroxymethyltransferase 2
SLC19A1	solute carrier family 19 (folate transporter), member 1
SNP	single nucleotide polymorphism
STR	short (simple) tandem repeat
TCN2	transcobalamin II
tHcy	total homocysteine
THF	tetrahydrofolate
TRDMT1	tRNA aspartic acid methyltransferase 1
TS	thymidylate synthetase
TYMS	thymidylate synthetase
UCP2	uncoupling protein-2
VNTR	variable number tandem repeat

CHAPTER 1

General introduction, objectives, and outline

1.1 Homocysteine is a key intermediate in one-carbon metabolism

Homocysteine is an intermediate sulphur-containing amino acid in one-carbon metabolism. One-carbon metabolism, which comprises the folate cycle and methionine pathway, regulates the transfer of carbon groups and is related to essential physiologic processes. These include formation of purines and thymidine for DNA and RNA synthesis, methylation of DNA, RNA, lipids and proteins, and regulation of oxidative stress. Homocysteine is formed in cells from the essential amino acid methionine via methyl transfer reactions (Figure 1.1). Homocysteine can be removed from the cell via (1) remethylation to methionine, (2) irreversible transsulfuration to cysteine (which is limited to cells of the liver, kidney, pancreas and gastro-intestinal tract), or (3) release of excess intracellular homocysteine into plasma. The intracellular concentration of homocysteine is regulated via these processes that require the action of several enzymes and cofactors. Approximately 70% of plasma homocysteine is disulphide bound to proteins. The remaining 30% is bound to cysteine or homocysteine to form homocysteine-cysteine or homocysteine-homocysteine (or homocystine) mixed disulfides. Only a very small portion is present as unbound, free homocysteine. The term plasma total homocysteine (tHcy) refers to the sum of all homocysteine species in plasma. Values for plasma tHcy in fasting state between 5 and 15 µmol/L are considered normal. Arbitrarily, moderate, intermediate and severe hyperhomocysteinemia (Hhcy) are classified as concentrations of >15-30 µmol/L, >30-100 µmol/L, and >100 µmol/L, respectively[1]. An example of the distribution of plasma tHcy concentrations in the general population, with its typical long tail toward higher plasma tHcy concentrations, is given in Figure 1.2.

1.2 Observational studies indicate associations between increased plasma tHcy and several chronic diseases but mechanism is unclear

Homocystinuria is a rare, autosomal recessive disorder, first described in the 1960s[2,3]. In general, it presents at young age and is characterized by severe Hhcy and extensive urinary excretion of homocystine. The clinical features of homocystinuria due to cystathione ß-synthase deficiency, the most common cause, include dislocated ocular lenses, mental retardation, osteoporosis and skeletal deformities, and vascular disease[4]. The presence of early-onset vascular disease in homocystinuria patients led to the hypothesis of a relation between mildly elevated concentrations of plasma tHcy and vascular damage[5,6].

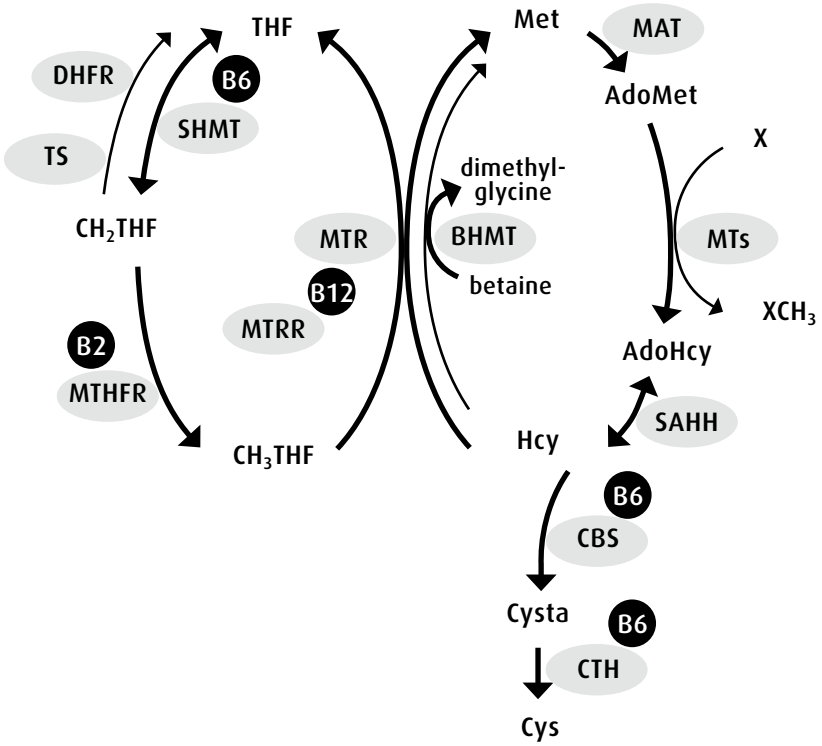

Figure 1.1 *Schematic overview of homocysteine metabolism*

Homocysteine (Hcy) is formed out of S-adenosylhomocysteine (AdoHcy), product from S-adenosylmethionine (AdoMet) dependent methyltransferase reactions. In all mammalian cells, Hcy can be remethylated to methionine (Met) by action of methionine synthase (MTR), an enzyme that uses vitamin B_{12} as cofactor and 5-methyltetrahydrofolate (CH_3THF) as methyldonor. Alternatively, in the liver and kidney, Hcy can be remethylated to Met in a reaction catalyzed by betaine-homocysteine S-methyltransferase (BHMT) and betaine as methyldonor. Hcy can be transsulfurated to cysteine (Cys) in an irreversible reaction of which the first step is catalyzed by cystathionine-beta-synthase (CBS) and in which vitamin B_6 serves as cofactor. If the intracellular Hcy concentration cannot be regulated effectively, excess Hcy will be transported into plasma.

AdoHcy: S-adenosylhomocysteine; AdoMet: S-adenosylmethionine; B2: vitamin B_2 (riboflavin); B6: vitamin B_6 (pyridoxine); B12: vitamin B_{12} (cobalamin); BHMT: betaine-homocysteine S-methyltransferase; CBS: cystathionine-beta-synthase; CH_2THF: 5,10-methylenetetrahydrofolate; CH_3THF: 5-methyltetrahydrofolate; Cys: cysteine; Cysta: cystathionine; CTH: cystathionase; DHFR: dihydrofolate reductase; Hcy: homocysteine; MAT: methionine adenosyltransferase; Met: methionine; MTs: AdoMet-dependent methyltransferases; MTHFR: methylenetetrahydrofolate reductase; MTR: methionine synthase; MTRR: methionine synthase reductase; SAHH: S-adenosylhomocysteine hydrolase; SHMT: serine hydroxymethyltransferase; THF: tetrahydrofolate; TS: thymidylate synthetase; X: methylacceptor.

Indeed, in 1976 Wilcken and Wilcken[7] showed abnormal homocysteine metabolism after methionine loading in patients with coronary artery disease. This first report was followed by many retrospective and prospective observational studies. These studies have been described in review articles and showed an overall positive association between plasma tHcy concentration and vascular disease (coronary heart disease, cerebrovascular disease, peripheral vascular diseases, and venous thrombosis)[8-13]. A meta-analysis in 2002 estimated that a reduction of 3 µmol/L plasma tHcy is associated with 16%, 25%, and 24% reduction for ischemic heart disease, deep vein thrombosis, and stroke, respectively[10]. Somewhat weaker effects were found in prospective and Mendelian randomization studies (see section 1.4)[9,14-16]. Randomized clinical trials to study the effect of homocysteine-lowering B-vitamin therapy on (secondary prevention of) mortality and morbidity from vascular disease were initiated a few years ago[17,18]. At the start of the current PhD study, the results of these randomized clinical trials were not yet published. The first publications and meta-analyses have shown inconclusive results with regard to effects on vascular disease risk and mortality[19-22] and have led to doubt about a causal and modifiable role for plasma tHcy in vascular disease pathology.

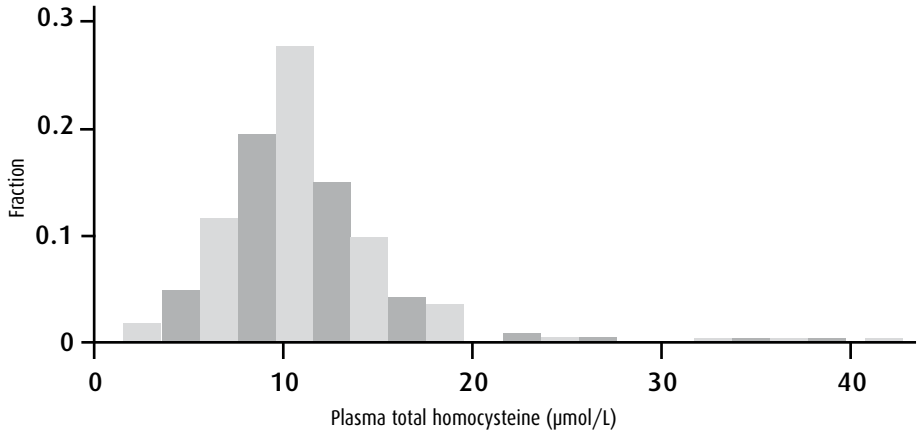

Figure 1.2 *Distribution of fasting plasma total homocysteine concentration in a population-based series of 461 Caucasian individuals (population is described in Chapters 3-6).*

In addition to vascular disease, several observational studies have proposed a role for increased plasma tHcy in the aetiology of neurological dysfunction in the elderly[23], Alzheimer's disease[24], schizophrenia[25], mood disorders[26], osteoporosis[27], and cancer[28,29].

Elevated concentrations of plasma tHcy have also been linked to increased risk for pregnancy complications and congenital abnormalities[30-32], including neural tube

defects (NTDs)[33-36]. The observation of decreased concentrations of folate in blood and diets of mothers of NTD-affected offspring triggered the initiation of randomized clinical trials for maternal periconceptional folic acid supplementation in the nineteen-eighties and early nineties. These trials showed that the occurrence and recurrence of NTD is reduced by 50-85%[37,38] and led to the general advice of periconceptional folic acid supplementation and to folic acid fortification in the United States.

The exact mechanism by which plasma tHcy concentration is related to disease is unknown and is likely to vary for different diseases. Numerous causal mechanisms, which call for downstream pathological effects of increased plasma tHcy, have been proposed. For vascular disease, these include promotion of endothelial dysfunction[39-41], oxidative stress[41], protein modification[41,42], and impairment of methylation ability[43]. The protective effect of folic acid supplementation on NTD risk clearly indicates a role for folate-related metabolism in NTD aetiology. However, the underlying mechanism of this protective effect and the inability to prevent NTDs in a substantial number of families is not yet fully understood. The increased maternal concentrations of plasma tHcy in NTD-affected pregnancies, which are inherently related to decreased folate concentrations, and the tHcy-lowering effect of folic acid supplementation suggested a role for homocysteine accumulation. Other proposed NTD mechanisms include maternal immunological responses that influence folate transport, impaired methylation rate (of DNA, RNA, lipids, proteins), and impaired nucleotide biosynthesis[36,44].

1.3 Elucidation of genetic aetiology of plasma tHcy-associated multifactorial diseases is important but difficult

The plasma tHcy-related diseases mentioned above (e.g. cardiovascular disease, venous thrombosis, NTDs) are, with exception of some rare monogenic forms, complex multifactorial diseases. Features of multifactorial diseases include the involvement of environmental as well as genetic risk factors, 'threshold' inheritance (presence of multiple genetic susceptibility variants is required), aetiological heterogeneity (phenotypically similar patients may have different underlying risk profiles), and presence of gene-gene (epistatic) and gene-environment interactions. In addition, the frequencies and penetrances of the involved genetic variants are likely to range from very small to moderate[45-47]. Elucidation of the genetic background of these complex diseases may aid in the understanding of the disease aetiology, the development of diagnostic and prognostic tools, and allow the identification of modifiable risk factors for the development of preventive and therapeutic measures. However, the abovementioned complexities make the identification and characterization of these genetic determinants for multifactorial diseases a challenging task[45-47].

1.4 Focus on plasma tHcy as genetically determined intermediate phenotype is easier and promotes elucidation of aetiology of multifactorial diseases

On the path between genetic variation and environmental exposures on one hand and discrete disease outcomes on the other, intermediate phenotypes can be defined[45,48]. Intermediate phenotypes can be viewed as measurable variables that may be related to underlying pathogenic mechanisms which are positioned more closely to elementary genetic and/or environmental disease susceptibility factors (Figure 1.3). Intermediate phenotypes that are influenced by genetic variations are also called endophenotypes[49]. Due to their intermediate location in disease aetiology the number of involved genes and genetic complexity is expected to be less which will make the genetic analysis of intermediates easier compared to that of the ultimate disease phenotype itself. In addition, if quantitative, the increased power of analysis of continuous compared to dichotomous disease traits can be exploited. Also, larger effect sizes can be expected[45,48]. An intermediate phenotype will be most useful in the search for genetic determinants of the associated complex diseases when it is a strong causal factor for disease and a sensitive indicator of the underlying pathogenic process, has a high heritability and shows a strong genetic correlation with the disease of interest, has a relatively simple genetic aetiology with no or limited aetiological (genetic) heterogeneity, and can be measured easily, precisely and reliably. Ideally, it is modifiable by environmental exposures (e.g. nutrition or drug) and hence offers direct target points for disease therapy.

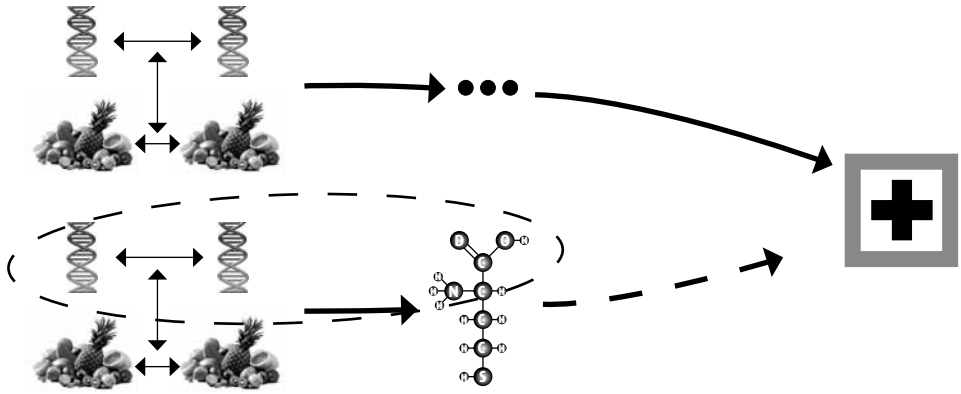

Figure 1.3 *Plasma tHcy concentration as intermediate phenotype for a multifactorial disease. Genetic and environmental factors, that may be interacting, underlie plasma tHcy concentrations and may also lead to disease. In addition, genetic and environmental factors that are unrelated to plasma tHcy are involved in disease aetiology. Elucidation of the genetic aetiology of plasma tHcy is less complex than that of the disease itself.*

Genetic determinants of intermediate phenotypes can be assessed for association to the disease outcome and provide direct indications for involved disease mechanisms. The associations between genetic variants and phenotype are generally insensitive for confounding by environmental factors due to the randomization that occurs during meiosis. Hence, the existence of randomly allocated genetic variants that predispose for an intermediate phenotype can be used to evaluate the effect of this intermediate phenotype on disease risk in an observational setting that is analogous to a randomized experimental design and allows causal inference. This concept was first introduced by Katan[50] and is known as 'Mendelian randomization'. The additional uses and limitations of Mendelian randomization have been described extensively by Davey Smith and Ebrahim[51]. Uses include the ability to study lifetime effects and the absence of "reversed causation" for fixed genetic variants. Limitations include the need for a reliable association between a genetic variant and the environmental risk factor under study, and the potential of confounding due to either population stratification or pleiotropic effects (a single genetic variant influences multiple phenotypes). In addition, very large sample sizes are required if the minor allele frequency (MAF) of the studied genetic variant and/or the effect on the intermediate phenotype is small.

The reported associations between plasma tHcy and multifactorial disease risk render this trait as a potentially useful intermediate one-carbon metabolism-related phenotype that can be modified by nutrition. The fact that several diseases are accompanied by elevated plasma tHcy concentrations, a heritable trait (see section 1.5), may imply the presence of shared underlying genetic factors. Indeed, Souto and colleagues[52] have estimated an age and sex adjusted genetic correlation between susceptibility to thrombosis and plasma tHcy of 0.652 which indicates that to a large extent the same genetic effects impact on both thrombosis and plasma tHcy. This may also hold for the other associated disorders. Hence, in the last decade or so, several groups have searched for genetic determinants of plasma tHcy concentrations.

1.5 Genetic underpinnings of plasma tHcy are not clear yet

Plasma tHcy concentrations may be elevated by physiologic and lifestyle factors, including dietary deficiencies of folate, vitamin B_{12} (cobalamin), vitamin B_6 (pyridoxal phosphate), and vitamin B_2 (riboflavin), old age, male sex, smoking, impaired renal function, and the use of certain drugs[53]. In addition, genetic causes affect the variation in plasma tHcy observed in the population. Several twin and family studies in healthy and diseased populations have evaluated the heritability (the part of the observed phenotypic variation of a trait or disease in a population that can be attributed to genetic variation within the population[54]) of plasma tHcy. The plasma tHcy concentrations in these studies were often analyzed with and without adjustment for a (sub)set

of plasma tHcy associated factors including age, sex, B-vitamin intake, folate, vitamin B_2, vitamin B_{12} and creatinin concentrations, smoking, alcohol consumption, and 677C>T genotype in the methylenetetrahydrofolate reductase gene (*MTHFR*). All but one[55] showed a significant contribution of variation in genes to the observed variation in plasma tHcy with reported heritability estimates varying from 28% to 63%[56-59].

The severely elevated plasma tHcy concentrations seen in homocystinuria patients are due to genetically determined deficiencies of enzymes in the homocysteine metabolism, including deficiency of methionine synthase and methylenetetrahydrofolate reductase but mostly cystathionine ß-synthase deficiency[4]. The mutations that cause these enzyme deficiencies, however, are very rare and hardly contribute to the heritability of plasma tHcy in the general population. Other, more common, DNA variants are presumed to be responsible for the heritable variation in plasma tHcy in the population. The knowledge about genetic causes of homocystinuria and homocysteine metabolism has been used in an attempt to efficiently select genes that have a high prior probability of influencing plasma tHcy concentrations, so called candidate genes.

In 2005, Gellekink et al.[60] reviewed epidemiological studies that assessed the association between candidate DNA variants and plasma tHcy concentrations. Table 1.1 is adapted from this review and reflects the status of research into genetic determinants for plasma tHcy at the start of this PhD project. Genetic research has focused on a limited number of key genes in folate and homocysteine metabolism and on single locus effects of non-synonymous DNA variants that were discovered by direct sequencing of coding regions. For some of the evaluated DNA variants, an effect on gene expression or functional activity has been demonstrated.

In 1995 the negative influence of the *MTHFR*677C>T variant on MTHFR enzyme activity and the resulting increase of plasma tHcy was described[61]. After that, numerous studies have replicated the association between the DNA variant and increased plasma tHcy. Approximately 12% of Caucasian individuals have the *MTHFR*677TT genotype and show ~50% of normal MTHFR activity and 2.5 µmol/L higher plasma tHcy than those with the *MTHFR*677CC genotype. Some have shown that the effect of this variant on plasma tHcy concentrations is dependent on folate and vitamin B_2 status[53]. Dependency of the effect of *MTHFR*677C>T on the 31 base-pair variable number of tandem repeats (VNTR) in cystathionine ß-synthase gene (*CBS*) has also been indicated[62]. In addition, interactions among *MTHFR*677C>T and 2576A>G in 5-methyltetrahydrofolate-homocysteine methyltransferase (*MTR*) and 66A>G in 5-methyltetrahydrofolate-homocysteine methyltransferase reductase (*MTRR*) have been reported[63]. Most other genetic determinants that have been evaluated for association to plasma tHcy have shown contradicting results among different studies. Also, the number of studies into these potential candidate variants is small and the estimated effects on plasma tHcy concentrations were often relatively small[60].

Table 1.1 *Genetic variants studied for their effect on plasma total homocysteine relating to non-fortified populations. Adapted with permission from Gellekink H. et al.*[60]

Gene symbol	Gene name	Chromosomal location	SNP ID	DNA variant	Amino acid change	Allele frequency	Reported effect on tHcy (Mutant vs Wild Type)
AHCY	S-adenosylhomocysteine hydrolase	20cen-q13.1	-	-34C>T	-	0.02 (T)	±
AHCY	S-adenosylhomocysteine hydrolase	20cen-q13.1	rs13043752	112C>T	Arg38Trp	0.03 (T)	No effect
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	-	595G>A	Gly199Ser	0.01 (A)	No effect
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	-	1218G>T	Glu406His	0.01 (T)	No effect
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	rs3733890	716G>A	Arg239Gln	0.22-0.31 (A)	No effect
CBS	cystathionine-beta-synthase	21q22.3	rs234706	699C>T	-	~0.36 (T)	No effect
CBS	cystathionine-beta-synthase	21q22.3	-	1444_1467+7(16_21) (31bp VNTR)	-	~0.77 (18x rpt)	+10% (18/18 vs 17/17)
CBS	cystathionine-beta-synthase	21q22.3	-	-5697(GT)10-20	-	0.67 (16x rpt)	No effect
CBS	cystathionine-beta-synthase	21q22.3	-	844_845ins(68bp)	-	~0.09 (ins)	0 to -23% (ns[1])
CBS	cystathionine-beta-synthase	21q22.3	rs1801181	1080C>T	-	~0.36 (T)	No effect
COMT	catechol-O-methyltransferase	22q11.21	rs4680	324G>A	Val108Met	0.48-0.75 (G)	-20%
CTH	cystathionase	1p31.1	rs1021737	1364G>T	Ser403Ile	0.29 (T)	+17%
FOLH1	glutamate carboxypeptidase II	11p11.2	-	1561C>T	His475Tyr	~0.06 (T)	-9% (ns)
MTHFD	methylenetetrahydrofolate dehydrogenase	14q24	rs2236225	2011G>A	Arg653Gln	0.40-0.45 (A)	No effect
MTHFR	5,10-methylenetetrahydrofolate reductase	1p36.3	rs1801133	677C>T	Ala222Val	0.30-0.40 (T)	+14 to +70%
MTHFR	5,10-methylenetetrahydrofolate reductase	1p36.3	rs1801131	1298A>C	Glu429Ala	~0.30 (C)	No effect

MTR	5-methyltetrahydrofolate-homocysteine methyltransferase	1q43	rs1805087	2756A>G	Asp919Gly	~0.20 (G)	0 to -20% (ns)
MTRR	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	5p15.31	rs1801394	66A>G	Ile22Met	0.46-0.59 (G)	0 to +10%
SHMT1	serine hydroxymethyltransferase 1	17p11.2	rs1979277	1420C>T	Leu474Phe	~0.30 (T)	±
SHMT2	serine hydroxymethyltransferase 2	12q13.2	-	1721_1722insTCTT	-	0.02 (del)	No effect
SLC19A1	solute carrier family 19 (folate transporter), member 1	21q22.3	rs1051266	80G>A	His27Arg	0.38-0.51 (A)	0 to +11% (ns)
TCN2	transcobalamin II	22q12.2	rs11557600	280G>A	Gly94Ser	0.01 (A)	No effect
TCN2	transcobalamin II	22q12.2	rs1801198	776C>G	Arg259Pro	0.35-0.47 (G)	0 to +15%
TCN2	transcobalamin II	22q12.2	rs9621049	1043C>T	Ser348Phe	0.11-0.17 (T)	No effect
TCN2	transcobalamin II	22q12.2	rs4820889	1196G>A	Arg399Gln	~0.02 (A)	No effect
TCN2	transcobalamin II	22q12.2	rs4820889	67A>G	Ile23Val	~0.13 (G)	-35% (ns)
TYMS	thymidylate synthetase	18p11.32	-	-220(28bp)2-4	-	0.17-0.47 (2x rpt)	No effect
TYMS	thymidylate synthetase	18p11.32	-	1494_1499delTTAAAG	-	0.36 (del)	No effect

[1] ns: nonsignificant

Several other studies have shown that the contribution of *MTHFR*677C>T and other DNA variants to variation in plasma tHcy is small to moderate and that other, yet unidentified, genetic factors remain to be identified[56,59,63] (especially for post-load plasma tHcy[57]); only one study suggested that the *MTHFR* locus is responsible for almost all the genetic variation in plasma tHcy[58]. This argues for a search for additional (single locus and multi-locus) genetic factors that contribute to plasma tHcy variance. Preferably, this search includes exploration of non-coding regions in candidate genes, evaluation of candidate genes that have so far not been studied, study of gene-gene interactions, and application of genome-wide mapping techniques[60].

1.6 Plasma tHcy-related genetic aetiology of NTD and venous thrombosis also unresolved

The presence of a genetic component in NTD aetiology has been emphasized by several studies[36,64-66]. Genes can contribute to congenital disease risk directly via the offspring genotype, with or without differential effects for maternal or paternal origin of the risk allele (parent-of-origin or imprinting effects). In addition, the maternal genotype can contribute to disease risk in offspring by influencing the intrauterine environment to which the developing offspring is exposed. A role for maternal genetic effects in NTD aetiology has been suggested[66]. The elucidation of genetic causes of NTD has been the focus of our research group for several years. Triggered by the NTD risk reduction due to periconceptional folic acid supplementation, genetic research has concentrated on a number of candidate genes in folate and homocysteine metabolism[36,65]. An extensive review on genetic variation related to homocysteine and folate metabolism and NTD risk has been written by van der Linden et al. in 2006[65]. In short, in 1995 *MTHFR*677C>T was identified by our group as first genetic NTD risk factor for mothers as well as offspring[67]. After that, a number of DNA variants in coding regions of genes related to folate and homocysteine metabolism have been explored for their association to NTDs in offspring or mothers. Generally, strong associations were not found but several variants may be associated to NTDs (e.g. maternal *MTRR*A66G). Research including variants other than the obvious candidate variants in folate and homocysteine metabolism, multi-locus effects, and more elaborate examination of maternal and offspring effects, may lead to identification of new genetic determinants of NTD[36,44,65].

In 1995, our group was one of the first to describe the association between plasma tHcy concentrations and thrombosis risk[68]. A causal role for plasma tHcy in thrombosis has been supported by some, but not all studies[22,69-71]. Recently, a role for low methionine concentrations in thrombosis aetiology was suggested[72]. The contribution of genetic variation to thrombosis liability has been estimated to ~55% in two studies[52,73].

The genetic correlation of 0.652 between plasma tHcy and liability to thrombosis estimated by Souto et al. [52] underlines the presence of shared genetic determinants. A recent meta-analysis by den Heijer et al.[13] demonstrated a positive association between *MTHFR*677C>T, an established determinant of plasma tHcy, and venous thrombosis although not all studies showed a significant result[13,74]. Also other homocysteine metabolism-related genetic variants have been evaluated for association to thrombosis risk in small scale studies. The 68 base pair insertion in *CBS* (*CBS*844ins68) did not show significant association in most studies[75-77] but one study did report a protective effect[78]. Inconsistent results regarding influence of combined presence of *CBS*844ins68 and *MTHFR*677C>T polymorphisms and thrombosis risk have been described[77-79]. Negative association results for *MTR*2576A>G[80-82], *MTRR*66A>G[75,82], 776C>G in the transcobalamin gene (*TCN2*)[83], and polymorphisms in S-adenosylhomocysteine hydrolase (*AHCY*), thymidylate synthetase (*TYMS*), reduced folate carrier (*RFC1* or *SLC19A1*) and AICAR transformylase/inosine monophosphate (*IMP*) cyclohydrolase (*ATIC*)[84,85] have been reported. Additional epidemiological studies into genetic variations that influence plasma tHcy as well as thrombosis risk are warranted to decipher the disease mechanisms underlying the association between plasma tHcy and thrombosis.

1.7 Alternative genetic epidemiological study designs and analyses may promote elucidation of genetic background of plasma tHcy and related phenotypes

In this thesis, the search for and characterization of genetic determinants of plasma tHcy concentration and venous thrombosis and NTD risk will involve a genetic epidemiological approach. Genetic epidemiology is the study of the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations[86]. Many similarities between concepts and methods used in classical epidemiology and the subspecialty of genetic epidemiology exist. However, the latter makes use of specific concepts and methodological tools to accommodate for the more or less predictable transmission from one generation to the next of the genetic determinants under study. Roughly, the process of genetic epidemiology involves establishment of a genetic component to the disease trait of interest, clarification of the underlying genetic inheritance model, localisation of the susceptibility genes, and identification and characterization of the causal genetic variants that underlie the disease trait[87].

As outlined in the sections above, epidemiological analysis of genetic determinants of plasma tHcy and associated disease phenotypes focused on direct association studies for potentially functional DNA variants in coding regions of a limited number of key genes in the folate cycle and homocysteine metabolism. This approach entails a high prior probability of actually genotyping the biologically relevant functional variant.

However, these hypothesis-driven candidate gene association studies rely on existing knowledge in the selection of genes and DNA variants of interest. Important DNA variants may be overlooked using this approach. Increase of the number of measured candidate genes and DNA variants can maximize the probability of genotyping the relevant variants. In addition, alternative approaches can be applied: linkage analysis and indirect association studies.

Linkage analysis allows for unbiased evaluation of the complete genome or genomic regions for DNA variants that influence disease traits. Linkage analysis involves the evaluation of cosegregation of polymorphic, non-functional DNA variants with known location (markers) with disease in families. Cosegregation indicates a lack of recombination and, hence, proximity of the marker and the disease locus. The ability to genotype sets of hundreds of markers spread throughout the human genome has contributed to the successful identification of many genes underlying monogenic and simple Mendelian diseases in the last two decades of the 20th century. The identified linkage region usually contains many genes and has to be scanned and evaluated in more detail to ultimately identify the specific genetic variant that causes the linkage signal[88]. Unfortunately, this linkage approach is not powerful for complex multifactorial disease traits; the influence of multiple genetic determinants with small marginal effects requires large number of families in which the disease trait segregates to obtain adequate power to map disease loci. Therefore, the focus of genetic epidemiology has shifted from linkage to association studies[89,90].

Association studies were facilitated by the development of rapid and cheaper genotyping techniques and the completion of the draft of the human genome sequence in 2003 and the subsequent increase in knowledge on DNA variants across the human genome[91]. One can distinguish between direct and indirect association studies. In the first, putative causal variants are genotyped, relying on existing knowledge about candidate DNA variants. Indirect association studies aim at measurement of causal variants via genotyping of DNA variants that are in linkage disequilibrium (LD) with these unmeasured, causal variants. The power to identify a causal variant is dependent on the strength of the LD measure r^2 that exists between the measured variants and the unmeasured causal locus. Optimal utilization of this indirect approach depends on knowledge of LD in the human genome. At the end of 2002 the International Haplotype Mapping (HapMap) project was initiated with the goal of constructing genome-wide maps of LD patterns by measuring a vast number of SNPs in four populations to facilitate the efficient execution of LD mapping[92]. The first and second phases of this project have been finished in 2005 and 2007, respectively[93,94] and the generated data are freely available via the World Wide Web (http://www.hapmap.org/). Hence, nowadays indirect association studies offer an alternative to the direct association approaches that have been used in the past and can be applied to candidate genes

or genomic regions to efficiently search for additional genetic variants for plasma tHcy and related disease phenotypes.

The statistical analysis of genetic association studies for plasma tHcy and related disease phenotypes generally involved a single DNA variant (single locus analysis). Analysis of multiple DNA variants simultaneously (i.e. multi-locus analysis) based on haplotypes (combinations of linked alleles at different loci on the same chromosomal background) can however be advantageous in indirect association studies where the r^2 between a haplotype and the unmeasured causal variant is stronger than that between the single DNA variants and the unmeasured causal variant or when allelic interactions are present. The analysis of combinations of genotypes that may be located in the same or different genes allows the study of multi-locus genotypic effects and interactions and increases the chance to detect causal DNA variants with small marginal effects but larger interactive effects[92]. Interaction analyses using traditional parametric regression analysis are hampered by problems including sparse data and multiple testing. However, a variety of advanced non-parametric techniques for analyzing multi-locus effects have been developed and described[95]. These have however not been widely applied yet and information on their performance is limited. More knowledge on the strengths and weaknesses may promote their correct application to real data, also for the analysis of genetic determinants of plasma tHcy.

The importance of distinguishing maternal from offspring genotypic effects in congenital disorders like NTD has been touched upon in the previous section. The population-based case-mother/control-mother design and the family-based case-parent triad design allow for this distinction. In the popular population-based case-control design association is based on evaluation of differences in (genotype or allele) frequencies of the DNA variant between case and control series. Associations may arise due to true direct or indirect association, chance, or confounding (due to other factors than LD). As stated earlier in section 1.3, associations between DNA variants and a disease trait are generally insensitive to confounding. Genetic case-control studies may be confounded, however, if the case and control samples are selected from a population that exists of several subpopulations with different allele frequencies of the genetic variant under study and differences in baseline disease risk. The extent to which population stratification may bias associations in populations of European origins, is still unclear. Several techniques to deal with confounding due to population stratification have been described and include selection of cases and controls from homogeneous outbred populations, division of the association test statistics of interest by the median test statistic of simultaneously measured genetic markers that are unrelated to the disease trait (genomic control), and the use of family-based association studies.

In the analysis of family-based designs (e.g. case-parent triads, nuclear families) the transmission rates of genotypes or alleles from parents to affected individuals are com-

pared with the untransmitted genotypes or alleles of those transmitted to unaffected family members. Hence, the control sample is inherently matched to the case sample with regard to genetic background and no bias from population stratification can arise. A disadvantage of these family-based studies is that ascertainment of study samples may be more difficult and that power is limited compared to equally-sized population-based case-control studies[92]. Hybrid designs, that entail genotyping of case-parent-triads and unrelated controls or unrelated control-parent pairs have recently been proposed and evaluated[96-98]. They were shown to exploit the advantages of both the population-based and family-based design. However, alternative hybrid designs can be constructed as well. Evaluation of these additional designs may enhance the toolbox for genetic epidemiological studies into genetic determinants of congenital disorders like NTDs.

1.8 Objectives and outline of this thesis

The main objective **(part 1)** of this thesis was to identify DNA variants that influence plasma total homocysteine concentrations and related diseases by use of genetic epidemiological studies. In this thesis we focused on neural tube defects and venous thrombosis. A second objective **(part 2)** was to develop and evaluate genetic epidemiological tools that support the main objective of this thesis and enhance the identification and characterization of genetic risk factors in multifactorial traits in general.

Part 1: Genetic determinants of homocysteine and related diseases
Chapter 2 describes a genome-wide linkage study that allows localisation of quantitative trait loci (QTLs) that influence plasma tHcy concentrations without the need to predefine candidate gene variants of interest. **Chapters 3 and 4** describe extensive candidate-gene association studies for plasma tHcy in which we measured DNA variation in genes that are involved in homocysteine and folate metabolism. In **chapter 3** we analysed the association between 79 variants in 40 genes and homocysteine, folate and methionine concentrations using a SNP-microarray approach. These 79 variants included mutations and polymorphisms that were known to change the function or regulation of a gene product and had already shown association with a disease but also contained synonymous and intronic SNPs that may indirectly measure untyped causal DNA variants via LD. In **chapter 4** we re-examined the association between 36 DNA variants in 19 genes and plasma tHcy. The single locus associations for these candidate variants were already reported previously. In the current study we applied haplotype and multi-locus genotype analyses to allow for the detection of haplotype effects and genotype interactions. In **chapters 5 and 6** we performed candidate-gene association studies for plasma tHcy *and* neural tube defects and venous thrombosis.

This enabled the simultaneous evaluation of the influence of the DNA variants on plasma tHcy concentration as well as the disease phenotype and hence the assessment of involved disease mechanisms. **Chapter 5** presents a case-control study on 45 folate-related genes and 87 DNA variants for spina bifida, the most common type of NTD, based on the same SNP-microarray as used in chapter 3. The DNA variants that were most strongly associated with spina bifida were assessed for association to homocysteine, folate and cobalamine concentrations in the control population. Unpublished preliminary results indicated a role for a 45 base pair deletion/insertion polymorphism in the uncoupling protein-2 gene (*UCP2*), a potential regulator of mitochondrial reactive oxygen species production and hence involved in oxidative stress, and plasma tHcy. In **chapter 6** the relation between this variant and plasma tHcy as well as recurrent venous thrombosis is evaluated in a case-control setting.

Part 2: Genetic epidemiological designs and analyses
Chapter 7 presents a new hybrid design that augments collection of case-parent triad and control-mother dyad genotype data to estimate offspring as well as maternal genetic effects in disease causation, an analysis that is especially relevant in congenital and early-onset disorders like NTDs. Its performance compared to existing alternative designs is evaluated. **Chapter 8** describes the evaluation of four existing techniques of analysis that can be applied to study multi-locus interactions in case-control studies: traditional parametric logistic regression, sum statistics, logic regression, and the multifactor dimensionality reduction method.

A general discussion of the studies that are presented in this thesis is given in **chapter 9**.

References

1. Ueland PM, Refsum H, Stabler SP, Malinow MR, Andersson A, Allen RH. Total homocysteine in plasma or serum: methods and clinical applications. Clin Chem. 1993;39:1764-1779.
2. Carson NAJ, Neill DW. Metabolic abnormalities detected in a survey of mentally backward individuals in Northern Ireland. Arch Dis Child. 1962:37;505-513.
3. Gerritsen T, Vaughn JG, Weisman HA. The identification of homocysteine in the urine. Biochem Biophys Res Commun. 1962; 9:493.
4. Mudd SH, Skovby F, Levy HL, Pettigrew KD, Wilcken B, Pyeritz RE, Andria G, Boers GH, Bromberg IL, Cerone R, Fowler B, Gröbe H, Schmidt H, Schweitzer L. The natural history of homocystinuria due to cystathionine beta-synthase deficiency. Am J Hum Genet. 1985; 37:1-31.

5. McCully KS. Vascular pathology of homocysteinemia: implications for the pathogenesis of arteriosclerosis. Am J Pathol. 1969;56:111-128.
6. McCully KS, Wilson RB. Homocysteine theory of arteriosclerosis. Atherosclerosis. 1975;22:215-227.
7. Wilcken DEL, Wilcken B. The pathogenesis of coronary artery disease – a possible role for methionine metabolism. J Clin Invest. 1976;57:1079-1082.
8. Boushey CJ, Beresford SA, Omenn GS, Motulsky AG. A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. Probable benefits of increasing folic acid intakes. JAMA. 1995;274:1049-1057.
9. Ueland PM, Refsum H, Beresford SA, Vollset SE. The controversy over homocysteine and cardiovascular risk. Am J Clin Nutr. 2000;72:324-332.
10. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. BMJ. 2002;325:1202.
11. Homocysteine Studies Collaboration. Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA. 2002;288:2015-2022.
12. Klerk M, Verhoef P, Clarke R, Blom HJ, Kok FJ, Schouten EG. MTHFR Studies Collaboration Group. MTHFR 677C-->T polymorphism and risk of coronary heart disease: a meta-analysis. JAMA. 2002;288:2023-2031.
13. den Heijer M, Lewington S, Clarke R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost. 2005;3:292-299.
14. Brattström L, Wilcken DE. Homocysteine and cardiovascular disease: cause or effect? Am J Clin Nutr. 2000;72:315-323.
15. Christen WG, Ajani UA, Glynn RJ, Hennekens CH. Blood levels of homocysteine and increased risks of cardiovascular disease. Causal or casual? Arch Intern Med. 2000;160:422-434.
16. Lewis SJ, Ebrahim S, Davey Smith G. Meta-analysis of MTHFR 677C->T polymorphism and coronary heart disease: does totality of evidence support causal role for homocysteine and preventive potential of folate? BMJ. 2005;331:1053.
17. Clarke R. Homocysteine-lowering trials for prevention of heart disease and stroke. Semin Vasc Med. 2005;5:215-222.
18. B-Vitamin Treatment Trialists' Collaboration. Homocysteine-lowering trials for prevention of cardiovascular events: a review of the design and power of the large randomized trials. Am Heart J. 2006;151:282-287.
19. Bazzano LA, Reynolds K, Holder KN, He J. Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials. JAMA. 2006;296:2720-2726.
20. Wald DS, Wald NJ, Morris JK, Law M. Folic acid, homocysteine, and cardiovascular disease: judging causality in the face of inconclusive trial evidence. BMJ. 2006;333:1114-1147.

21. Wang X, Qin X, Demirtas H, Li J, Mao G, Huo Y, Sun N, Liu L, Xu X. Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. Lancet. 2007;369:1876-1882.
22. den Heijer M, Willems HP, Blom HJ, Gerrits WB, Cattaneo M, Eichinger S, Rosendaal FR, Bos GM. Homocysteine lowering by B vitamins and the secondary prevention of deep vein thrombosis and pulmonary embolism: A randomized, placebo-controlled, double-blind trial. Blood. 2007;109:139-144.
23. Morris MS. Folate, homocysteine, and neurological function. Nutr Clin Care. 2002;5:124-132.
24. Morris MS. Homocysteine and Alzheimer's disease. Lancet Neurol. 2003;2:425-428.
25. Muntjewerff JW, Kahn RS, Blom HJ, den Heijer M. Homocysteine, methylenetetrahydrofolate reductase and risk of schizophrenia: a meta-analysis. Mol Psychiatry. 2006; 11:143-149.
26. Bottiglieri T. Homocysteine and folate metabolism in depression. Prog Neuropsychopharmacol Biol Psychiatry. 2005;29:1103-1112.
27. Morris MS, Jacques PF, Selhub J. Relation between homocysteine and B-vitamin status indicators and bone mineral density in older Americans. Bone. 2005;37:234-242.
28. Choi SW, Mason JB. Folate and carcinogenesis: an integrated scheme. J Nutr. 2000;130:129-132.
29. Powers HJ. Interaction among folate, riboflavin, genotype, and cancer, with reference to colorectal and cervical cancer. J Nutr. 2005;135:2960S-2966S.
30. Ray JG, Laskin CA. Folic acid and homocyst(e)ine metabolic defects and the risk of placental abruption, pre-eclampsia and spontaneous pregnancy loss: A systematic review. Placenta. 1999;20:519-529.
31. Nelen WL. Hyperhomocysteinaemia and human reproduction. Clin Chem Lab Med. 2001;39:758-763.
32. Refsum H. Folate, vitamin B12 and homocysteine in relation to birth defects and pregnancy outcome. Br J Nutr. 2001;85 Suppl 2:S109-113.
33. Mills JL, McPartlin JM, Kirke PN, Lee YJ, Conley MR, Weir DG, Scott JM. Homocysteine metabolism in pregnancies complicated by neural-tube defects. Lancet. 1995;345:149-151.
34. van der Put NM, van Straaten HW, Trijbels FJ, Blom HJ. Folate, homocysteine and neural tube defects: an overview. Exp Biol Med (Maywood). 2001;226:243-270.
35. Kirke PN, Mills JL, Scott JM. Homocysteine metabolism in pregnancies complicated by neural tube defects. Nutrition. 1997;13:994-995.
36. Blom HJ, Shaw GM, den Heijer M, Finnell RH. Neural tube defects and folate: case far from closed. Nat Rev Neurosci. 2006;7:724-731.
37. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. Lancet. 1991;338:131-137.
38. Czeizel AE, Dudás I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation. N Engl J Med. 1992;327:1832-1835.

39. Faraci FM. Hyperhomocysteinemia: a million ways to lose control. Arterioscler Thromb Vasc Biol. 2003;23:371-373.
40. Undas A, Brozek J, Szczeklik A. Homocysteine and thrombosis: from basic science to clinical evidence. Thromb Haemost. 2005;94:907-915.
41. Lentz SR. Mechanisms of homocysteine-induced atherothrombosis. J Thromb Haemost. 2005;3:1646-1654.
42. Jacobsen DW, Catanescu O, Dibello PM, Barbato JC. Molecular targeting by homocysteine: a mechanism for vascular pathogenesis. Clin Chem Lab Med. 2005;43:1076-1083.
43. James SJ, Melnyk S, Pogribna M, Pogribny IP, Caudill MA. Elevation in S-adenosylhomocysteine and DNA hypomethylation: potential epigenetic mechanism for homocysteine-related pathology. J Nutr. 2002;132:2361S-2366S.
44. Beaudin AE, Stover PJ. Folate-mediated one-carbon metabolism and neural tube defects: balancing genome synthesis and gene expression. Birth Defects Res C Embryo Today. 2007;81:183-203.
45. Schork NJ. Genetics of complex disease: approaches, problems, and solutions. Am J Respir Crit Care Med. 1997;156:S103-S109.
46. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet. 2001;17:502-510.
47. Sing CF, Stengård JH, Kardia SL. Genes, environment, and cardiovascular disease. Arterioscler Thromb Vasc Biol. 2003;23:1190-1196.
48. Goldman D, Ducci F. Deconstruction of vulnerability to complex diseases: enhanced effect sizes and power of intermediate phenotypes. ScientificWorldJournal. 2007;7:124-130.
49. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry. 2003;160:636-645.
50. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet. 1986;i:507–508
51. Davey Smith G, Ebrahim S, What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? BMJ. 2005;330:1076-1079.
52. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet. 2000;67:1452-1459.
53. Refsum H, David Smith A, Ueland PM, Nexo E, Clarke R, McPartlin J, Johnston C, Engbaek F, Schneede J, McPartlin C, Scott JM, Facts and recommendations about total homocysteine determinations: an expert opinion. Clinical Chemistry. 2004;50:3-32.
54. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era – concepts and misconceptions. Nature Rev Genet. 2008;9:255-266.
55. Cesari M, Burlina AB, Narkiewicz K, Sartori MT, Sacchetto A, Rossi GP. Are fasting plasma homocyst(e)ine levels heritable? A study of normotensive twins. J Investig Med. 2000;48:351-358.

56. Jee SH, Song KS, Shim WH, Kim HK, Suh I, Park JY, Won SY, Beaty TH. Major gene evidence after MTHFR-segregation analysis of serum homocysteine in families of patients undergoing coronary arteriography. Hum Genet. 2002;111:128-135.
57. den Heijer M, Graafsma S, Lee SY, van Landeghem B, Kluijtmans L, Verhoef P, Beaty TH, Blom H. Homocysteine levels--before and after methionine loading--in 51 Dutch families. Eur J Hum Genet. 2005;13:753-762.
58. Bathum L, Petersen I, Christiansen L, Konieczna A, Sørensen TI, Kyvik KO. Genetic and environmental influences on plasma homocysteine: results from a Danish twin study. Clin Chem. 2007;53:971-979.
59. Siva A, De Lange M, Clayton D, Monteith S, Spector T, Brown MJ. The heritability of plasma homocysteine, and the influence of genetic variation in the homocysteine methylation pathway. QJM. 2007;100:495-499.
60. Gellekink H, den Heijer M, Heil SG, Blom HJ. Genetic determinants of plasma total homocysteine. Semin Vasc Med. 2005;5:98-109.
61. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJH, den Heijer M, Kluijtmans LAJ, van den Heuvel LP, Rozen R, A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet. 1995;10:111-113.
62. Afman LA, Lievers KJ, Kluijtmans LA, Trijbels FJ, Blom HJ. Gene-gene interaction between the cystathionine beta-synthase 31 base pair variable number of tandem repeats and the methylenetetrahydrofolate reductase 677C > T polymorphism on homocysteine levels and risk for neural tube defects. Mol Genet Metab. 2003;78:211-215.
63. Kluijtmans LA, Young IS, Boreham CA, Murray L, McMaster D, McNulty H, Strain JJ, McPartlin J, Scott JM, Whitehead AS. Genetic and nutritional factors contributing to hyperhomocysteinemia in young adults. Blood. 2003;101:2483-2488.
64. Mitchell LE, Adzick NS, Melchionne J, Pasquariello PS, Sutton LN, Whitehead AS. Spina bifida. Lancet. 2004;364:1885-1895.
65. van der Linden IJ, Afman LA, Heil SG, Blom HJ. Genetic variation in genes of folate metabolism and neural-tube defect risk. Proc Nutr Soc. 2006;65:204-215.
66. Deak KL, Siegel DG, George TM, Gregory S, Ashley-Koch A, Speer MC; NTD Collaborative Group. Further evidence for a maternal genetic effect and a sex-influenced effect contributing to risk for human neural tube defects. Birth Defects Res A Clin Mol Teratol. 2008;82:662-669.
67. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet. 1995;346:1070-1071.
68. den Heijer M, Blom HJ, Gerrits WB, Rosendaal FR, Haak HL, Wijermans PW, Bos GM. Is hyperhomocysteinaemia a risk factor for recurrent venous thrombosis? Lancet. 1995;345:882-885.
69. Cattaneo M. Hyperhomocysteinemia and venous thromboembolism. Semin Thromb Hemost. 2006;32:716-723.

70. Ray JG, Kearon C, Yi Q, Sheridan P, Lonn E; Heart Outcomes Prevention Evaluation 2 (HOPE-2) Investigators. Homocysteine-lowering therapy and risk for venous thromboembolism: a randomized trial. Ann Intern Med. 2007;146:761-767.
71. Naess IA, Christiansen SC, Romundstad PR, Cannegieter SC, Blom HJ, Rosendaal FR, Hammerstrøm J. Prospective study of homocysteine and MTHFR 677TT genotype and risk for venous thrombosis in a general population--results from the HUNT 2 study. Br J Haematol. 2008;141:529-535.
72. Keijzer MB, den Heijer M, Borm GF, Blom HJ, Vollset SE, Hermus AR, Ueland PM. Low fasting methionine concentration as a novel risk factor for recurrent venous thrombosis. Thromb Haemost. 2006;96:492-497.
73. Heit JA, Phelps MA, Ward SA, Slusser JP, Petterson TM, De Andrade M. Familial segregation of venous thromboembolism. J Thromb Haemost. 2004;2:731-736.
74. Bezemer ID, Doggen CJ, Vos HL, Rosendaal FR. No association between the common MTHFR 677C->T polymorphism and venous thrombosis: results from the MEGA study. Arch Intern Med. 2007;167:497-501.
75. Akar N, Akar E, Misirlioglu M, Avcu F, Yalcin A, Cin S. Search for genetic factors favoring thrombosis in Turkish population. Thromb Res. 1998;92:79-82.
76. Franco R, Maffei F, Lourenco D, Piccinato C, Morelli V, Thomazini I, Zago M. The frequency of 844ins68 mutation in the cystathionine beta-synthase gene is not increased in patients with venous thrombosis. Haematologica. 1998;83:1006-1008.
77. de Franchis R, Fermo I, Mazzola G, Sebastio G, Di Minno G, Coppola A, Andria G, D'Angelo A. Contribution of the cystathionine ß-synthase gene (844ins68) polymorphism to the risk of early-onset and arterial occlusive disease and of fasting hyperhomocysteinemia. Thromb Haemost. 2000;84:576-582.
78. Grossmann R, Schwender S, Geisen U, Schambeck C, Merati G, Walter U. CBS 844ins68, MTHFR TT677 and EPCR 4031ins23 genotypes in patients with deep-vein thrombosis. Thrombosis Research. 2002;107:13-15.
79. Gaustadnes M Rudiger N, Rasmussen K, Ingerslev J. Intermediate and severe hyperhomocysteinemia with thrombosis: a study of genetic determinants. Thromb Haemost. 2000;83:554-558.
80. Ray G, Langman AJ, Vermeulen MJ, Evrovski J, Yeo EL, Cole DE. Genetic University of Toronto Thrombophilia Study in Women (GUTTSI): genetic and other risk factors for venous thromboembolism in women. Curr Control Trials Cardiovasc Med. 2001;2:141-149.
81. Salomon O, Rosenberg N, Zivelin A, Steinberg DM, Kornbrot N, Dardik R, Inbal A, Seligsohn U. Methionine synthase A2756G and methylenetetrahydrofolate reductase A1298C polymorphisms are not risk factors for idiopathic venous thromboembolism. Hematol J. 2001;2:38-41.
82. Gellekink H, Kluijtmans LA, Blom HJ, den Heijer M. Disturbed vitamin B12 metabolism, variation in homocysteine remethylation genes and recurrent venous thrombosis risk. In: Gellekink H: Molecular genetic analysis of hyperhomocysteinemia. Radboud University Nijmegen; 2007. pp 27-38.

83. Pereira AC, Lourenço DM, Maffei FH, Morelli VM, Rollo HA, Zago MA, Vannucchi H, Franco RF. A transcobalamin gene polymorphism and the risk of venous thrombosis. The BRATROS (Brazilian Thrombosis Study). Thromb Res. 2007;119:183-188.
84. Gellekink H, den Heijer M, Kluijtmans LA, Blom HJ. Effect of genetic variation in the human S-adenosylhomocysteine hydrolase gene on total homocysteine concentrations and risk of recurrent venous thrombosis. Eur J Hum Genet. 2004;12:942-948.
85. Gellekink H, Blom HJ, den Heijer M. Associations of common polymorphisms in the thymidylate synthase, reduced folate carrier and 5-aminoimidazole-4-carboxamide ribonucleotide transformylase/inosine monophosphate cyclohydrolase genes with folate and homocysteine levels and venous thrombosis risk. Clin Chem Lab Med. 2007;45:471-476.
86. Khoury MJ, Beaty TH, Cohen BH. Fundamentals of genetic epidemiology. 1993. New York: Oxford University Press.
87. Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. Lancet. 2005;366:941-951.
88. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228-237.
89. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516-1517.
90. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Genetics Rev. 2002;3:1-7.
91. Palmer L, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. Lancet. 2005;366:1223-1234.
92. Cordell HJ, Clayton DG. Genetic association studies. Lancet. 2005;366:1121-1131.
93. International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299:1320.
94. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851-861.
95. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet. 2003;4:701-709.
96. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet. 2005;76:592-608.
97. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet. 2004;12:964-970.
98. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet. 2005;77:627-636.

PART 1

Genetic determinants of homocysteine and related diseases

SHHM Vermeulen
GM van der Vleuten
J de Graaf
AR Hermus
HJ Blom
AFH Stalenhoef
M den Heijer

Published in: Journal of Thrombosis and Haemostasis 2006;4:1303–1307

CHAPTER 2

A genome-wide linkage scan for homocysteine levels suggests three regions of interest

Abstract

Background

An elevated plasma total homocysteine (tHcy) level is a risk factor for many clinical conditions, including vascular disease and venous thrombosis. The tHcy levels are partly determined by genetic factors. Extensive candidate gene studies have identified several genetic variants, including the MTHFR 677C>T, that influence tHcy levels, but so far only part of the genetic variation in tHcy can be explained.

Objective

In order to identify chromosomal regions that influence tHcy levels, a genome-wide linkage analysis was conducted.

Patients/methods

Our study population consisted of 13 pedigrees and 469 subjects with data on fasting plasma tHcy levels. A set of 377 markers covering the genome was genotyped in 275 subjects. The variance component linkage method (SOLAR version 2.1.3) was used for the two-point and multipoint linkage analyses.

Results

The heritability of the age- and sex-adjusted homocysteine levels was 44%. The multipoint linkage analysis identified one region with suggestive linkage on chromosome 16q (LOD score 1.76; nominal P=0.0024). Weaker evidence of linkage was found for regions on chromosome 12q (LOD score 1.57; nominal P=0.0036) and chromosome 13q (LOD score 1.52; nominal P=0.0041).

Conclusions

In our families the plasma tHcy level was highly heritable. The multipoint linkage analysis identified three regions that showed weak to suggestive linkage to tHcy levels.

Introduction

Homocysteine is a sulfur-containing intermediate in the methionine metabolism. An elevated level of plasma total homocysteine (tHcy), or hyperhomocysteinemia, is associated with many clinical conditions, including vascular disease[1–3], stroke[1,4], venous thrombosis[3,5], congenital abnormalities [6] and Alzheimer's disease[7].

A substantial part of the variation in plasma tHcy levels can be contributed to genetic factors; heritability estimates vary from 8 to 47%[8–10]. To date, the search for genetic determinants of plasma tHcy levels has been focusing on the candidate gene approach, in which genes are selected on basis of our current knowledge of the homocysteine metabolism. Genetic variants within these candidate genes have been evaluated for their association with plasma tHcy levels, mostly in population-based association studies. The most important genetic determinant identified so far is the MTHFR 677C>T polymorphism. This single nucleotide polymorphism has been reported to explain 4–9% of the variation in plasma tHcy levels[9–12]. Other genetic variants have been studied, but they showed weak and inconsistent effects[13]. So, despite extensive candidate gene studies, it seems that only little of the genetic variation of plasma tHcy levels can be explained.

Here, we report one of the first conducted genome-wide linkage scans for plasma tHcy levels. An advantage of this method is that it does not rely on the selection of candidate genes beforehand. Our objective was to find quantitative trait loci (QTLs) that contribute to variation in plasma tHcy levels, which may lead to the identification of genes that have not been found through the candidate gene approach.

Subjects and methods

Study population

The study population is part of a large family study on familial combined hyperlipidemia (FCH) in which families were recruited in 1994 and followed up in 1999. This population has been described previously[14,15]. Briefly, ascertainment of the families took place through probands who were diagnosed as having FCH based on the presence of high cholesterol and/or high triglyceride levels. A family was only included if a multiple type of hyperlipidemia was found in first-degree relatives and the proband or a first-degree relative presented premature cardiovascular disease before the age of 60 years. Subjects had to be 12 years or older to be included, and all individuals were Caucasian. In total, 50 families comprising 1036 individuals entered the study. The ethical committee of the Radboud University Nijmegen Medical Centre approved the study protocol.

Thirteen informative families consisting of 482 subjects were selected for the genome scan. These families varied in size from 12 to 114 individuals and spanned at least three generations. Genotyping has been performed in 275 persons who were selected based on their positions in the pedigree and expected informativity. Selection of the pedigrees and 275 persons was performed using the SLINK simulation software[16,17].

Plasma homocysteine measurement

Blood samples of individuals in a subpopulation of the FCH study, including the 13 genome scan pedigrees, were collected by venipuncture after a 4-week withdrawal period of lipid-lowering medication and an overnight fast, and were put on ice immediately. Total fasting homocysteine concentration in plasma was measured by an automated high-performance liquid chromatography method with reverse phase separation and fluorescence detection[18]. The plasma tHcy level was successfully determined in 469 of the 482 subjects.

Genotyping

DNA was obtained from peripheral blood lymphocytes by a standard method[19]. Genotyping of 377 markers in the autosomal genes at an average 10 cM density and with an average marker heterozygosity of 0.75 was performed by the Marshfield Mammalian Genotyping Service (Marshfield, WI, USA). The markers primarily originated from screening set 9 and in part from screening set 10. Sex-averaged marker maps were obtained from the Marshfield Center of Medical Genetics (http://research.marshfieldclinic.org/genetics/). The MTHFR 677C>T polymorphism was tested by polymerase-chain reaction in 333 subjects, essentially according to the method of Frosst et al.[20]. Discrepancies in genotyping were checked using the PedCheck program[21] and inconsistent marker genotypes were set to missing (0.5%).

Statistical analysis

Quantitative genetic analysis and two-point and multipoint linkage analysis were performed using the variance components method as implemented in the SOLAR package (version 2.1.3)[22]. The variance components method assumes multivariate normality of the trait and is vulnerable to a high kurtosis[23]. In order to obtain more normally distributed values, homocysteine levels were logarithmically transformed and z-scores were calculated. The homocysteine values of two outliers were reduced to the next less extreme value. Prior to heritability calculations and linkage analysis, adjustments for age and sex were made using the regression modeling option. Allele frequencies of the markers were estimated through maximum likelihood techniques and were used to impute missing genotype data and to create identical by descent (IBD) files. The program HOMO that is implemented in the SOLAR package was used to assess genetic heterogeneity. Simulation techniques were applied to calculate empirical point wise

P-values of observed LOD scores. As pedigrees were selected through FCH probands and no association between tHcy levels and FCH exists in these pedigrees[18], no ascertainment correction was made. Using simulation techniques in SOLAR, we calculated our power to detect a QTL with an effect size of 25%, that is, a locus that explains 25% of variation in plasma tHcy levels, based on the number of observations with both information on plasma tHcy level and genome scan data. The power to find the QTL with a LOD score of 2, 1.5 and 1 or higher was 41%, 57% and 75%, respectively. Therefore, LOD scores of 1 or higher were reported here.

Results

The mean age of the 469 subjects with homocysteine measurements who were available for the quantitative genetic analysis was 43.3 years (range: 13–88 years). The percentages of males (45%) and females (55%) were comparable. After correction of the two outliers with tHcy values of 1.9 and 46.6 μmol L^{-1}, the geometric mean fasting tHcy level was 11.0 μmol L^{-1} with a range of 4.1–34.2 μmol L^{-1}. The heritability of the unadjusted log transformed plasma tHcy levels was 40.7%. After adjustment for age and sex the heritability increased to 44.4% ($P < 0.0001$).

Two hundred and sixty-four subjects had both homocysteine measurements and genome scan data and were available for the linkage analysis. Their mean age was 48.5 years (range 14– 88 years), 47% was male and 53% was female, and the geometric mean tHcy level in this subgroup was 11.1 μmol L^{-1} (range 4.3–34.2 μmol L^{-1}).

MTHFR 677C>T polymorphism

The MTHFR 677C>T polymorphism was genotyped in 333 subjects with tHcy measurements. Geometric means of plasma homocysteine levels (95% confidence intervals; CI) for the MTHFR677 genotype groups were 9.9 (9.4–10.4), 10.7 (10.2–11.2) and 12.1 (10.2–14.3) μmol L^{-1} for the CC, CT and TT group, respectively. The MTHFR 677C>T polymorphism was associated with plasma tHcy levels and accounted for 3.0% of the variation in tHcy levels in this study population.

Multipoint linkage analysis

The results of the multipoint genome scan for the adjusted plasma tHcy levels are shown in Fig. 2.1. Three areas with a LOD score exceeding 1 were observed. The region with the highest score of 1.76 (nominal P=0.0024) was located on chromosome 16q12 between markers D16S3253 and GATA138C05 and a peak location of 75 cM. The linkage peak was very broad with a 1-LOD support interval (interval in which the LOD score is within 1.0 of its maximum) of 33 cM till 90 cM. Furthermore, tentative indications of linkage to regions on chromosome 12q14 (LOD score 1.57; nominal P=0.0036) and

chromosome 13q31 (LOD score 1.52; nominal *P*=0.0041) were identified. The linkage regions on chromosome 12 and 13 were also very broad with 1-LOD support intervals of 40–92 cM and 68–100 cM and peak locations of 83 cM and 87 cM, respectively. The reported LOD score on chromosome 16 could be found by chance approximately one time every genome scan, and can be interpreted as suggestive linkage evidence[24]. The other LOD scores did not meet the criterion for suggestive linkage. No genetic heterogeneity was observed for the three regions.

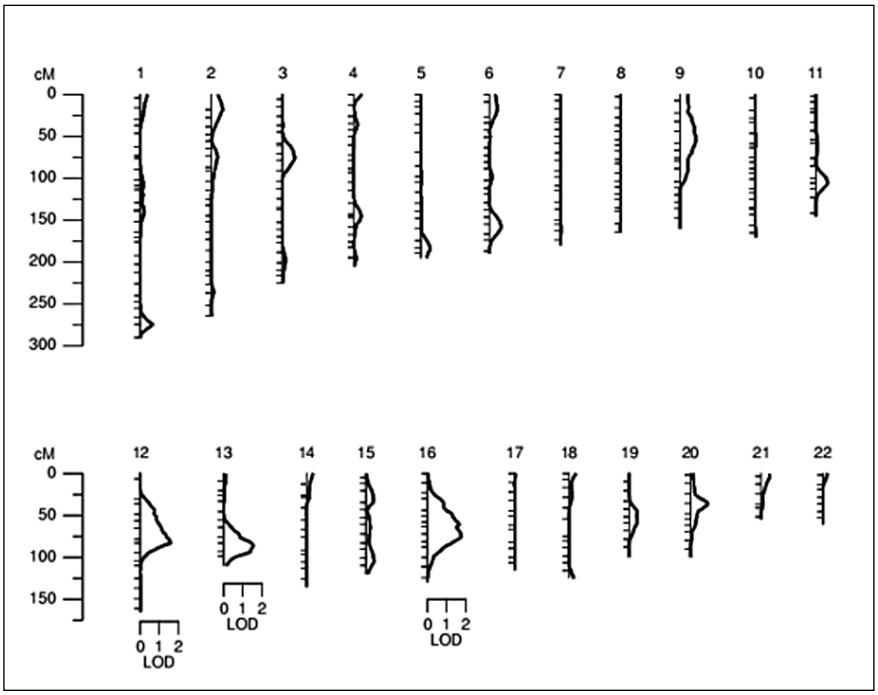

Fig. 2.1 *Results from the genome-wide multipoint linkage scan for the autosomal chromosomes. The hatch marks on the left side of the chromosome indicate the locations of the genotyped markers. The bold line indicates the LOD score for the chromosomal locus.*

The multipoint linkage analysis is sensitive to errors in marker maps. Therefore, two-point linkage analysis was conducted. The three multipoint peaks were also found in the two-point analysis with LOD scores of 2.0 (D12S1052, 83.2 cM), 1.4 (D13S796, 93.5 cM) and 1.3 (D16S3253, 71.8 cM) for chromosomes 12, 13 and 16, respectively. Other linkage peaks in the two-point analysis were located on chromosome 1 (D1S1609, 274.5 cM, LOD score 1.5), chromosome 3 (D3S1768, 61.5 cM, LOD score 1.6), and chromosome 20 (GATA129B03, 35.5 cM, LOD score 1.1) but these were not observed in the multipoint analysis.

Discussion

We conducted a genome-wide linkage analysis for plasma tHcy levels in an FCH population consisting of 13 extended pedigrees comprising 469 subjects with homocysteine measurements and 275 subjects with genome scan data. The heritability of age- and sex-adjusted plasma tHcy levels was 44%, indicating that in our population additive genetic effects determined a substantial part of the variation in homocysteine levels. Three regions on chromosome 12, 13 and 16 with weak to suggestive linkage to QTLs for tHcy were identified.

Souto and colleagues recently reported results of the first genome scan for plasma tHcy levels in Spanish families in the GAIT Study [25]. In this study, 12 families were ascertained through a proband with thrombophilia, and nine were recruited without regard to phenotype. Significant linkage evidence (LOD score 3.1) was found for chromosome 11q23 and suggestive linkage (LOD score 1.7) for chromosome 10q22. The nicotinamide N-methyltransferase (NNMT) gene, a positional candidate gene located in the 11q23 linkage region, showed significant association with plasma tHcy levels. In our study, we could not confirm these linkage results. However, a small linkage peak was observed proximal of the NNMT location with a peak LOD score of 0.61 at 105 cM. Differences or non-replication of the results of the linkage scan in the Spanish families and ours may have arisen because of (genetic) heterogeneity between the populations or to relatively low power of both studies to identify QTLs with small effects.

Homocysteine level can be regarded as a multifactorial trait; environmental factors and several genes with different effect sizes are involved and gene–environment and gene–gene interactions can play a role. It is common knowledge that moderate size linkage studies can lack power to identify small effect QTLs in complex traits with high LOD scores and that, as in every study with (multiple) testing, there is a chance of false positive signals that has to be reckoned with. This calls for a careful interpretation of the linkage results that, possibly together with information from other studies and sources, could lead to new hypotheses regarding chromosomal regions or genes involved in the trait of interest. Simulations showed that in our study we had 41% power to detect a QTL that explained 25% in tHcy variation or more with a LOD score of 2. Our linkage study was not powerful enough to provide conclusive linkage evidence for QTLs with relatively small effect sizes.

Jee et al.[10] suggest the presence of a major gene for homocysteine in addition to MTHFR that may explain as much as 37% of the variation in adjusted homocysteine levels, based on results of a family study with probands who underwent elective coronary arteriography. The presence of an additional major gene independent of the MTHFR 677C>T could not be confirmed by a segregation analysis we performed in 51 families ascertained through hyperhomocysteinemic probands[9]. The MTHFR 677C>T polymorphism was measured in 333 subjects in this study; it was associated with

plasma tHcy levels and accounted for ~3% of plasma tHcy variation in our families. With the heritability estimate of 44% in mind, these findings indicate that other genes besides the MTHFR can in theory play a major role in determining plasma tHcy levels. However, our results do not support the presence of a major QTL with a very large effect, as indicated in the article by Jee et al.[10].

Although the linkage regions we found were very broad, we scanned the three regions on chromosome 16, 12 and 13 for biologically plausible candidate genes that could explain the observed linkage peaks with help of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway). The 1-LOD support interval around the peak on chromosome 16q was mapped to a region of ~60 cM and contains numerous genes. We could not select an obvious candidate gene. Also, we did not find positional candidate genes in the linkage region on chromosome 13.

The region on chromosome 12 contains the serine hydroxy-methyltransferase 2 (SHMT2) gene that is located on chromosome 12q13.2 and encodes the mitochondrial isoform of this methyltransferase. This enzyme catalyzes the reversible reaction of serine and tetrahydrofolate into glycine and 5,10-methylene tetrahydrofolate. Theoretically, deleterious genetic variants within this gene could lead to elevation of tHcy levels. A few years ago, we already investigated this gene in 70 patients with spina bifida but no significant association with tHcy levels was found[26]. As far as we know, no other association studies for variants in this gene and plasma tHcy have been performed. It is therefore questionable if genetic variants in this gene could be responsible for the observed linkage signal.

Furthermore, the methyltransferase-like gene 1 (METTL1) that is assigned to chromosome location 12q13 lies within the 1-LOD-support interval. This small gene contains a highly conserved S-adenosylmethionine-binding domain and therefore probably encodes a methyltransferase[27]. METTL1 could be viewed as a possible positional candidate gene. It has not been studied for its relation with tHcy levels yet, but further studies to explore its association could be warranted.

Other known candidate genes involved in the homocysteine metabolism were not found in the linkage analysis. The MTHFR gene is the most important known genetic determinant of plasma tHcy level and maps to chromosome 1p36.3. We did not find evidence of linkage for this region; a LOD score of ~0.2 was observed. With help of simulation studies we calculated an expected LOD score (ELOD) of ~0.2 for a QTL with the effect size of 3% as we observed for the MTHFR 677C>T polymorphism in this study. Not finding a high linkage peak for the MTHFR gene in this study could be expected because of its low impact on tHcy levels in this population.

Overall, one region with suggestive linkage evidence and two other potentially interesting regions were identified. Two candidate genes that may be responsible for the linkage peak on chromosome 12 are SHMT2 and METTL1 but their relation to the linkage peak is very speculative. Investment in fine mapping of the linkage regions

could narrow the 1-LOD support intervals and refine the location of the putative genetic determinants. Confirmation of the linkage regions or confirmation of possible involvement of the genomic regions in homocysteine levels in other studies is warranted before further steps are taken.

Acknowledgements

This study was supported by the Netherlands Heart Foundation, Grant 2002B68. Martin den Heijer is supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO). Genotyping was kindly performed by the Mammalian Genotyping Service at the Marshfield Center of Medical Genetics. We would like to thank M. J. Veerkamp and S. J. Bredie who were crucial in recruiting the families. We are very grateful to all the families who participated in the FCH study.

References

1. Homocysteine Studies Collaboration. Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA. 2002;288:2015–2022.
2. Klerk M, Verhoef P, Clarke R, Blom HJ, Kok FJ, Schouten EG. MTHFR 677C>T polymorphism and risk of coronary heart disease: a meta-analysis. JAMA. 2002;288:2023–2031.
3. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. BMJ. 2002;325:1202.
4. Casas JP, Bautista LE, Smeeth L, Sharma P, Hingorani AD. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. Lancet. 2005;365: 224–232.
5. den Heijer M, Lewington S, Clarke R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost. 2005;3:292–299.
6. Nelen WL. Hyperhomocysteinaemia and human reproduction. Clin Chem Lab Med. 2001;39:758–763.
7. Morris MS. Homocysteine and Alzheimer's disease. Lancet Neurol. 2003;2:425–428.
8. Cesari M, Burlina AB, Narkiewicz K, Sartori MT, Sacchetto A, Rossi GP. Are fasting plasma homocyst(e)ine levels heritable? A study of normotensive twins. J Invest Med. 2000;48:351-358.
9. den Heijer M, Graafsma S, Lee SY, van Landeghem B, Kluijtmans L, Verhoef P, Beaty TH, Blom H. Homocysteine levels – before and after methionine loading – in 51 Dutch families. Eur J Hum Genet. 2005;13:753–762.
10. Jee SH, Song KS, Shim WH, Kim HK, Suh I, Park JY, Won SY, Beaty TH. Major gene evidence after MTHFR-segregation analysis of serum homocysteine in families of patients undergoing coronary arteriography. Hum Genet. 2002;111:128–135.

11. Gaughan DJ, Kluijtmans LA, Barbaux S, McMaster D, Young IS, Yarnell JW, Evans A, Whitehead AS. The methionine synthase reductase (MTRR) A66G polymorphism is a novel genetic determinant of plasma homocysteine concentrations. Atherosclerosis. 2001;157:451–456.
12. Kluijtmans LA, Young IS, Boreham CA, Murray L, McMaster D, McNulty H, Strain JJ, McPartlin J, Scott JM, Whitehead AS. Genetic and nutritional factors contributing to hyperhomocysteinemia in young adults. Blood. 2003;101:2483–2488.
13. Gellekink H, den Heijer M, Heil SG, Blom HJ. Genetic determinants of plasma total homocysteine. Semin Vasc Med. 2005;5:98–109.
14. Veerkamp MJ, de Graaf J, Bredie SJ, Hendriks JC, Demacker PN, Stalenhoef AF. Diagnosis of familial combined hyperlipidemia based on lipid phenotype expression in 32 families: results of a 5-year follow-up study. Arterioscler Thromb Vasc Biol. 2002;22:274–282.
15. Veerkamp MJ, de Graaf GJ, Hendriks JC, Demacker PN, Stalenhoef AF. Nomogram to diagnose familial combined hyperlipidemia on the basis of results of a 5-year follow-up study. Circulation. 2004;109:2980–2985.
16. Ott J. Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA. 1989;86:4175–4178.
17. Weeks DE, Ott J, Lathrop GM. SLINK: a general simulation program for linkage analysis. Am J Hum Genet. 1990;47:A204.
18. Veerkamp MJ, de Graaf J, den Heijer M, Blom HJ, Stalenhoef AF. Plasma homocysteine in subjects with familial combined hyperlipidemia. Atherosclerosis. 2003;166:111–117.
19. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16:1215.
20. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, den Heijer M, Kluijtmans LA, van den Heuvel LP, Rozen R. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet. 1995;10:111–113.
21. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet. 1998;63:259–266.
22. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998;62:1198–1211.
23. Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet. 1999;65:531–544.
24. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet. 1995;11:241–247.
25. Souto JC, Blanco-Vaca F, Soria JM, Buil A, Almasy L, Ordonez-Llanos J, Martin-Campos M, Lathrop M, Stone W, Blangero J, Fontcuberta J. A genomewide exploration suggests a new candidate gene at chromosome 11q23 as the major determinant of plasma homocysteine levels: results from the GAIT Project. Am J Hum Genet. 2005;76:925–933.

26. Heil SG, Van der Put NM, Waas ET, den Heijer M, Trijbels FJ, Blom HJ. Is mutated serine hydroxymethyltransferase (SHMT) involved in the etiology of neural tube defects? Mol Genet Metab. 2001;73:164–172.
27. Bahr A, Hankeln T, Fiedler T, Hegemann J, Schmidt ER. Molecular analysis of METTL1, a novel human methyltransferase-like gene with a high degree of phylogenetic conservation. Genomics. 1999;57:424–428.

Sita HHM Vermeulen
Barbara Franke
Marieke JH Coenen
Mascha MVAP Schijvenaars
Hans Scheffer
Per M Ueland
Henk J Blom
Martin den Heijer

Submitted

CHAPTER 3

Candidate-gene association study for one-carbon metabolism-related genes and folate, homocysteine, and methionine concentrations

Abstract

Concentrations of folate, homocysteine, and methionine have been associated with common disorders like neural tube defects, vascular disease, and cancer. Identification of genetic determinants of these one-carbon metabolism intermediates can be used as a tool to elucidate pathogenic mechanisms and identify modifiable risk factors. Previous studies have focussed on a limited number of key genes and have not led to complete understanding of the genetic underpinnings of these three metabolites. Therefore, we analyzed 79 DNA variants in 40 genes related to the folate cycle, remethylation, transmethylation, and transsulfuration for association with fasting serum folate, fasting and post-load plasma total homocysteine (tHcy), and fasting plasma methionine concentration in 190 population-based Caucasian subjects. SNP rs1801133 (*MTHFR*677C>T) was main determinant of serum folate (nominal p=0.002); other strong determinants included rs8101626 in *DNMT1* (nominal p=0.003) and rs1801394 (*MTRR*66A>G) (nominal p=0.011). *MTHFR*677C>T was also nominally associated to fasting and post-load plasma tHcy (nominal p=0.026 and 0.006, respectively). SNP rs2276598 (*DNMT*3A1523G>A) showed nominal association to fasting tHcy (nominal p=0.035). *CBS*844_845ins(68bp) was significant determinant of post-load tHcy (nominal p=0.0005; family-wise p=0.044) and explained 6.3% of its variance. A non-synonymous variant (rs672346) in *BHMT* (nominal p=0.014) and 3′ UTR rs1078004 in *ICMT* (nominal p=0.030) were related to methionine. Substantial overlap in strong DNA variants for fasting and post-load plasma tHcy was present. Only *MTHFR*677C>T showed strong association to more than one of the three metabolites. In conclusion, our study resulted in establishment of known and identification of new candidate DNA variants for folate, homocysteine, and methionine concentrations.

Introduction

One-carbon metabolism involves the transfer of carbon groups and is related to essential physiologic processes. These include formation of purines and thymidine for DNA and RNA synthesis, methylation of DNA, RNA, lipids and proteins, and regulation of oxidative stress (Figure 3.1). Concentrations of the one-carbon intermediates folate, homocysteine, and, to a lesser extent, methionine, have been associated to the development of numerous disorders including neural tube defects (NTDs), vascular disease, cancer and neurological disorders [12]. The mechanisms underlying the association between clinical conditions and these correlated metabolites are still unclear and have been proposed to be causal. This has been supported by the fact that periconceptional folic acid supplementation reduces the occurrence and recurrence of NTDs by 50-85%[3,4]. Also, a causal role has been indicated by numerous observational studies, including genetic association studies on the 677C>T single nucleotide polymorphism (SNP; rs1801133) in the methylenetetrahydrofolate reductase gene (*MTHFR*) that exploit the principle of Mendelian randomization[5], (e.g. [6-8]). However, contradicting observational findings have also been published (e.g. [9,10]), and recent secondary intervention trials with plasma tHcy-lowering therapy, including folic acid, were inconclusive with regard to effects on vascular disease risk and mortality[11-13].

Proper function of one-carbon metabolism and maintenance of adequate concentrations of its intermediate metabolites are dependent on genetic factors in addition to habitual dietary intake of folate and B-vitamins and biological factors like age and sex[14]. The contribution of genetic variability to the population interindividual variation of serum folate, reflecting short-term fluctuation in dietary intake and mainly present in the form of 5-methyl-THF, has been examined in one study; heritability was estimated to be 32% (after adjustment for age, sex and smoking)[15]. No heritability estimates for methionine concentrations have been published, to our knowledge. In contrast, several reports on the heritability of plasma tHcy have been published with estimates varying from 8% to 63%[16-20]. Notably, variation in post-methionine load (hereafter referred to as post-load) plasma tHcy is presumably under stronger influence of genetic variation than fasting state plasma tHcy[18].

The identification of DNA variants that influence the concentrations of disease-related one-carbon intermediates can increase our understanding of the underlying pathology of one-carbon metabolism-related diseases, and may lead to the identification of nutrition-gene interactions that contribute to disease risk. Although numerous genetic association studies for plasma tHcy and, to a lesser extent, folate, have been reported, the *MTHFR*677C>T is the only DNA variant for which the relation with decreased folate and increased plasma tHcy is established. Since its discovery in 1995[21] the association of this SNP, which affects MTHFR activity, and plasma tHcy and folate has been confirmed in multiple populations[22-26]. The study of other DNA variants in key genes

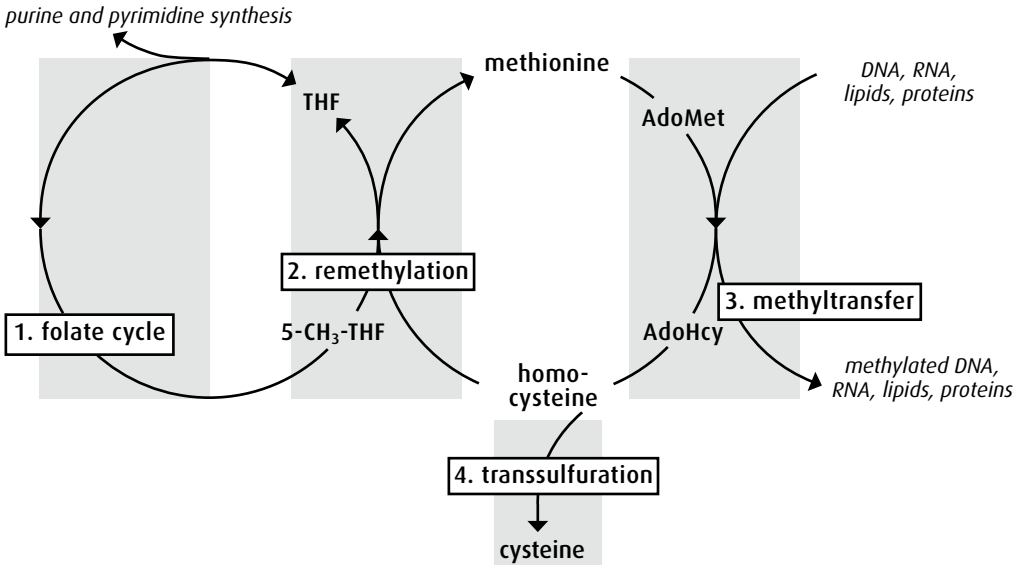

Figure 3.1 *Simplified representation of one-carbon metabolism with depiction of four metabolic processes (see Table 3.1; outlined and numbered), major metabolites (bold), and metabolic substrates and products (italics).*

One-carbon metabolism regulates the availability of methyl groups for essential physiologic processes. Activation of methionine by adenosine triphosphate (ATP) results in formation of S-adenosylmethionine (AdoMet), a universal methyl donor for methylation of deoxyribonucleic acid (DNA), ribonucleic acid (RNA), lipids and proteins. S-adenosylhomocysteine (AdoHcy), the demethylated product of these reactions, is the precursor of homocysteine. Homocysteine can be irreversibly transsulfurrated to cysteine and, eventually, glutathionine. The latter is involved in oxidative stress mechanisms. Homocysteine can also be remethylated to methionine. The methyl donor for this reaction is 5-methyltetrahydrofolate (5-CH_3THF). The remethylation reaction thus involves donation of a methyl group from the folate cycle to the methionine cycle. Tetrahydrofolate (THF), formed in this remethylation reaction, but also via other reactions, is substrate for purine and pyrimidine synthesis, the building blocks of DNA and RNA.

in one-carbon metabolism has not led to conclusive results and identification of additional major determinants[25,26]. However, the 31 base pairs (bp) variable number of tandem repeats (VNTR) polymorphism in *cystathionine beta-synthase* (*CBS*), 766G>C (rs1801198) in *transcobalamin II* (*TCN2*), 324G>A (rs4680) in *catechol-o-methyl-transferase* (*COMT*), and 1364G>T (rs1021737) in *cystathionine γ-lyase* (*CTH*), may be regarded as weaker determinants of (post-load or fasting) plasma tHcy that warrant further investigation[25]. Recently, a population-based study including 10,601 subjects investigated the associations of 12 one-carbon-related metabolites, including folate, homocysteine and methionine concentrations, with 13 candidate-gene DNA variants[24]. They did not find significant associations between methionine concentrations

and any of the investigated polymorphisms but did find associations with plasma tHcy and *MTHFR*677C>T, *MTHFR*1286A>C, 2756A>G in *5-methyltetrahydrofolate-homocysteine methyltransferase* (*MTR*), and the *CBS*844_845ins68 variant. In addition, associations for serum folate with *MTHFR*677C>T, *MTHFR*1286A>C, 1958G>A in *methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1* (*MTHFD1*), and 80G>A in *solute carrier family 19 (folate transporter), member 1* (*SLC19A1*, *RFC1*) were reported.

The aim of this study was to identify genetic determinants of folate, homocysteine and methionine concentrations. So far, the search for genetic determinants of folate and homocysteine has mainly focused on a restricted number of DNA variants in several relevant genes. In the present study, we have applied a custom SNP-microarray approach to allow simultaneous genotyping of a large number of DNA variants in genes related to one-carbon metabolism and that span the processes of folate metabolism, remethylation, methyl transfer, and transsulfuration. Seventy-nine DNA variants in 40 genes were analyzed for their association with fasting and post-methionine load plasma tHcy, fasting serum folate, and fasting plasma methionine in 190 Caucasian subjects.

Methods and materials

Study sample

One hundred and ninety unrelated subjects of self-reported Caucasian ethnicity out of a population-based sample comprising 500 individuals were included in the current study. These 500 individuals were ascertained via a general practice in The Hague in 1993 and have been described previously. The study protocol was approved by the ethics committee of the Leyenburg Hospital[27].

Metabolite measurements

Venous blood samples were collected after an overnight fast and 6 hours after an oral methionine loading (100 mg L-methionine per kg bodyweight dissolved in 200 ml orange juice). EDTA plasma was stored at –20°C, serum at –70°C. Serum folate was measured using the Dualcount SPNB (solid phase no boil) Radioassay (Diagnostic product Corporation, Los Angeles, USA). Plasma tHcy was measured according to the method described by Fiskerstrand et al.[28] which was slightly modified[29]. Concentration of methionine in plasma was measured using a modification of the method described by Holm et al.[30]. To account for oxidation of methionine during storage we calculated total methionine as the sum of methionine and methionine sulfoxide.

Selection of candidate genes and DNA variants and genotyping

The selection and genotyping of candidate genes and DNA variants has been described previously[31]. Briefly, we selected the variants based on literature and databases

(dbSNP, URL: http://www.ncbi.nlm.nih.gov/SNP/; SNPper at Chip Bioinformatics Tools, URL: http://snpper.chip.org/bio/) during 2002. For DNA variants selected from literature, we focused on those that were known to change the function or regulation of a gene product and those that had already shown association with (any) disease or condition. If such DNA variant could not be found in a gene, we searched the databases for SNPs. We preferably selected those that had been confirmed (e.g. by at least two submissions), had a known frequency in the Caucasian population and were potentially functional (due to location in splice sites, promoters or other regulatory regions, or by changing an amino acid). If these SNPs were not found in a gene, we selected SNPs within or near exons and in the promoter region to identify associations due to linkage disequilibrium. In genes in which functional polymorphisms were available, we did not select additional SNPs for analysis. In total, 50 genes and 154 DNA variants were selected for genotyping. Genotyping of most SNPs was carried out at ASPER Biotech (Tartu, Estonia) using arrayed primer extension (APEX) technology[32]. In case the design of an APEX assay for a SNP was not possible as well as for non-SNP variants, genotyping was carried out at the Departments of Human Genetics and Pediatrics in Nijmegen, The Netherlands. Assay conditions for these DNA variants are available from the corresponding author.

We included only those DNA variants in the statistical analysis that (1) had a minor allele frequency (MAF) higher than 0.05, (2) had less than 25% missing genotype data, and (3) were in Hardy-Weinberg equilibrium (HWE) (*P*-value>0.01). Seventy-nine DNA variants in 40 genes, classified according to five one-carbon metabolic processes, fulfilled these criteria (Table 3.1).

Statistical analysis

Tests for deviation from HWE were performed by means of Chi-square tests. Fisher's exact test was chosen when expected genotype counts dropped below five. For multi-allelic variants a Monte Carlo exact test was applied. The pair-wise linkage disequilibrium (LD) measure r^2 was estimated for di-allelic variants using Haploview (http://www.broad.mit.edu/haploview/haploview?q=mpg/haploview)[47]. Distributions of metabolite concentrations were skewed to the right and therefore logarithmically transformed. Residual metabolite values, adjusted for age and sex, and pair-wise Pearson correlation coefficients (r) were calculated using Stata version 9.1. Tests for association of the DNA variants with metabolite concentrations were performed using Plink version 1.02 (http://pngu.mgh.harvard.edu/purcell/plink/)[48]. In the main analysis we evaluated the association between the DNA variants and metabolites assuming an additive effect of allele dosage. Multi-allelic variants were recoded for all association analyses into di-allelic variants (Table 3.1). Nominal empirical p-values and family-wise empirical p-values (corrected for all performed tests for the metabolite) were generated using 2000 permutations.

Table 3.1 *Characteristics of DNA variants included in analysis.*

Gene symbol [A]	Gene name [A]	Metabolic process[B]	DNA variant[C]	Protein change	MAF[D] (%)	Nr. genotyped	Reference[E]
AHCY	S-adenosylhomocysteine hydrolase	3	rs1205366		14.0	190	
AHCY	S-adenosylhomocysteine hydrolase	3	rs819158		11.3	164	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs1868138		24.2	190	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs3796191	Leu254Pro	9.2	190	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs2305230		15.8	190	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs873696		38.2	190	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs2290053		49.2	190	
ALDH1L1	aldehyde dehydrogenase 1 family, member L7	1	rs1127717	Asp793Gly	19.5	190	31
AMT	aminomethyltransferase	3	rs10640		25.4	189	
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	1	rs2372536	Thr116Ser	31.2	181	34
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	1	rs1997059		33.1	189	
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	1	rs4673991		33.1	189	
BHMT	betaine-homocysteine methyltransferase	2	rs672346	Phe219Leu	7.1	189	
BHMT	betaine-homocysteine methyltransferase	2	rs651852		50.0	189	
BHMT	betaine-homocysteine methyltransferase	2	rs3733890	Arg239Gln	29.4	189	35
BHMT2	betaine-homocysteine methyltransferase 2	2	rs682985		42.3	187	
BHMT2	betaine-homocysteine methyltransferase 2	2	rs526264		41.5	188	
CBS	cystathionine-beta-synthase	4	g.14037(31bp)16-21 [F,G]			181	36
CBS	cystathionine-beta-synthase	4	844_845ins(68bp) [H]		8.7	183	37

Gene symbol [A]	Gene name [A]	Metabolic process[B]	DNA variant[C]	Protein change	MAF[D] (%)	Nr. genotyped	Reference[E]
COQ3	coenzyme Q3 homolog, methyltransferase	3	rs7755791		31.0	187	
CTH	cystathionase	4	rs1021737	Ser403Ile	30.0	190	
CTH	cystathionase	4	rs3767205		8.2	190	
CTH	cystathionase	4	rs501939		26.5	183	
CTH	cystathionase	4	rs490574		24.5	190	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	2	rs932640		7.9	177	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	2	rs1907362		7.9	190	31
CUBN	cubilin (intrinsic factor-cobalamin receptor)	2	rs1801231	Pro1559Ser	8.7	190	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	2	rs2271461		19.1	189	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	1	rs1801239	Ile2984Val	12.9	190	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	2	rs703075		29.2	190	
DHFR	dihydrofolate reductase	1	IVS1+261delACCTGGGCGGGACGCGCCA		40.1	182	38
DNMT1	DNA (cytosine-5-)-methyltransferase 1	3	rs2290684		49.7	189	
DNMT1	DNA (cytosine-5-)-methyltransferase 1	3	rs8101626		44.7	190	
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	3	rs2276598		19.0	190	
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	3	rs1465764		6.6	190	
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	3	rs2289195		36.5	152	
FOLR1	folate receptor 1 (adult)	1	rs1893007		5.3	188	
FOLR3	folate receptor 3	1	rs533207		11.7	188	
FOLR3	folate receptor 3	1	rs1802608	Leu193Phe	5.8	190	
FTCD	formiminotransferase cyclodeaminase	1	rs12774		28.7	190	
FTCD	formiminotransferase cyclodeaminase	1	rs7277617		21.8	190	
FTCD	formiminotransferase cyclodeaminase	1	rs4819205		6.4	187	
GART	phosphoribosylglycinamide formyltransferase	1	rs8971	Asp752Gly	21.1	190	
GGH	gamma-glutamyl hydrolase	1	rs719235		25.4	189	
GGH	gamma-glutamyl hydrolase	1	rs1031552		7.4	189	

ICMT	isoprenylcysteine carboxyl methyltransferase	3	rs1802353		5.8	190	
MAT1A	methionine adenosyltransferase I, alpha	3	rs2342812		49.7	190	
MAT2A	methionine adenosyltransferase II, alpha	3	rs1078004		49.5	189	
MGMT	O-6-methylguanine-DNA methyltransferase	3	rs12917	Leu84Phe	14.2	190	
MGMT	O-6-methylguanine-DNA methyltransferase	3	rs2020893	Glu30Lys	14.2	190	
MGMT	O-6-methylguanine-DNA methyltransferase	3	rs2308321	Ile143Val	12.6	190	
MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1	1	rs2236225	Arg653Gln	47.9	190	39
MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1	1	rs1950902	Lys134Arg	16.1	190	
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2	1	rs12196		20.8	190	
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2	1	rs1667627		42.5	186	
MTHFR	5,10-methylenetetrahydrofolate reductase	2	rs1801133	Ala222Val	31.1	185	40, 41
MTHFR	5,10-methylenetetrahydrofolate reductase	2	rs1801131	Ala429Glu	31.6	190	41
MTHFS	5,10-methenyltetrahydrofolate synthetase	1	rs2586183		46.2	182	
MTR	5-methyltetrahydrofolate-homocysteine methyltransferase	2	rs1805087	Asp919Gly	15.8	190	42
MTRR	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	2	rs1801394	Ile22Met	42.0	188	43
NNMT	nicotinamide N-methyltransferase	3	rs1941404		37.1	190	
NOS1	nitric oxide synthase 1	5	IVS14+267(AAT)8-16 [I]			145	
NOS1	nitric oxide synthase 1	5	4775(CA)14-22 [J]			168	
NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	5	rs2297518	Ser608Leu	15.7	185	
NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	5	IVS1-2660(CCTTT)8-18 [K]			167	
NOS3	nitric oxide synthase 3 (endothelial cell)	5	IVS4+245(GAAGTCTAGACCTGCTGCAGGGGTGAG)4-6 [L]			168	
NOS3	nitric oxide synthase 3 (endothelial cell)	5	rs1799983	Asp298Glu	31.1	190	44

Gene symbol [A]	Gene name [A]	Metabolic process[B]	DNA variant[C]	Protein change	MAF[D] (%)	Nr. genotyped	Reference[E]
PRMT1	protein arginine methyltransferase 1	3	rs975484		25.8	188	
RNMT	RNA (guanine-7-) methyltransferase	3	rs4797810		11.1	189	
SARDH	sarcosine dehydrogenase	1	rs573904		27.4	190	31
SARDH	sarcosine dehydrogenase	1	rs2073815	His489His	41.3	190	
SARDH	sarcosine dehydrogenase	1	rs2073817	Arg614His	37.4	190	
SARDH	sarcosine dehydrogenase	1	rs7854480		47.6	189	
SHMT1	serine hydroxymethyltransferase 1 (soluble)	1	rs1979277	Leu474Phe	31.1	190	45
SLC19A1	solute carrier family 19 (folate transporter), member 1	1	rs1051266	His27Arg	43.4	189	31, 34
TCN2	transcobalamin II	2	rs1801198	Arg259Pro	44.2	190	46
TRDMT1	tRNA aspartic acid methyltransferase 1	3	rs2295809		49.1	167	31
TYMS	thymidylate synthetase	1	-97(CCGCGCCACTTGGCCTGCCTCC-GTCCCG)2-4 [M]			180	34
TYMS	thymidylate synthetase	1	[CCGCGCCACTTGGCCTGCCTCCGTCCCG]>[CCGCGCCACTTCGCCTGCCTCCGTCCCG] [N]			176	

[A] approved symbol and name according to HUGO Gene Nomenclature Committee (http://www.genenames.org/; page last updated: July 16, 2007)
[B] 1: folate cycle; 2: remethylation; 3: AdoMet-dependent methyl transfer; 4: transsulfuration; 5: other
[C] rs number is given where available; otherwise, nomenclature according to den Dunnen and Antonarakis[47] is given
[D] minor allele frequency
[E] reference in which gene-metabolite association in this study population has been published before
[F] for heterogeneous repeat units see Lievers et al.[36]
[G] recoded into di-allelic variant; allele 1 = 18; allele 2 = 17 or 19 or 20
[H] 68 bp: CATCCAGGTGGGGTTTTGCTGGGCTTGAGCCCTGAAGCCGCGCCCTCTGCAGATCATTGGGGTGGAT
[I] recoded into di-allelic variant; allele 1 = less than 12 repeats; allele 2 = 12 or more repeats
[J] recoded into di-allelic variant; allele 1 = 17 repeats; allele 2 = more or less than 17 repeats
[K] recoded into di-allelic variant; allele 1 = 8 or 9 or 10 or 11 or 12 repeats; allele 2 = 13 or 14 or 15 or 16 or 17 repeats
[L] only 2 alleles present (4 or 5 repeats) in population
[M] only 2 alleles present (2 or 3 repeats) in population
[N] G>C transversion in the 3 repeat allele of the -97(CCGCGCCACTTGGCCTGCCTCCGTCCCG)2-4 in *TYMS*

In addition, we performed a multi-locus set-based analysis as implemented in Plink. The five sets of DNA variants were defined using the categorization of the genes in five metabolic processes as outlined in Figure 3.1 and Table 3.1. This set-test is based on calculating the average test statistic for the best DNA variant per set, for the best two variants per set, et cetera. Empirical p-values unadjusted for multiple testing, adjusted for all tests within each set, and adjusted for all tests in all sets were generated using 2000 permutations. We excluded DNA variants rs2372536 and rs1997059 in *amino-imidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase* (*ATIC*) and rs682985 in *betaine-homocysteine methyltransferase 2* (*BHMT2*) since they were strongly correlated with other DNA variants within these genes (see Results). We allowed a maximum of five DNA variants in each sum for each set. This set-based analysis allowed us to explore which variant or variants within each metabolic process are most strongly associated to the metabolites and permits comparison of the importance of the genetic variants within and between metabolites. In all analyses, two-sided p-values were calculated and values < 0.05 were considered statistically significant.

Results

DNA variants measured and analysed previously

Fifteen out of the 79 DNA variants have been measured and analysed for association to folate and/or plasma tHcy in (subsets of) the population-based sample, previously. Also, for four additional DNA variants we have recently evaluated the association to fasting RBC and serum folate and plasma tHcy in the 190 subjects used in the current study[31]. Table 3.1 indicates these 19 DNA variants and includes the references in which the associations were described. For completeness and additional analyses we have included these DNA variants also in the current study.

Metabolite concentrations

Of the 190 subjects included in this study, 40% was male and median (5^{th} - 95^{th} percentile) age was 43.1 (27.0 - 55.2) years. Median (5^{th} - 95^{th} percentile) fasting serum folate, plasma fasting and post-load tHcy, and methionine were 12.1 (5.7 – 30.1) nmol/L, 10.4 (5.2 – 18.4) µmol/L, 36.3 (23.1 – 77.0) µmol/L, and 23.7 (17.4 – 32.5) µmol/L, respectively. A strong correlation between fasting and post-load plasma tHcy (Pearson's r=0.652; p<0.0001) was observed. Moderate inverse correlations were found between fasting plasma tHcy and serum folate, and post-load plasma tHcy and serum folate (Pearson's r=-0.364, and -0.308, respectively (all $p \leq 0.0001$)). No significant correlations with plasma methionine were observed.

Table 3.2 *Allelic association results for 10 DNA variants that showed strongest association to fasting serum folate, fasting and post-load plasma total homocysteine (tHcy), and fasting plasma methionine.*

		fasting folate			fasting tHcy			postload tHcy			fasting methionine		
Gene	**DNA variant**	**nominal p**	**effect[A]**	**R^2 [B]**	**nominal p**	**effect**	**R^2**	**nominal p**	**effect**	**R^2**	**nominal p**	**effect**	**R^2**
Folate cycle													
ALDH1L1	rs1127717										0.066	-	1.8
ATIC	rs1997059				0.013	-	3.4	0.065	-	1.9			
ATIC	rs2372536				0.034	-	2.3						
ATIC	rs4673991				0.011	-	3.4	0.053	-	2			
ATIC	rs4673991												
FOLR1	rs1893007	0.028	+	2.6				0.085	-	1.6			
FTCD	rs4819205				0.048	-	2.3	0.014	-	3.3			
GART	rs8971										0.041	-	2.2
MTHFD1	rs1950902				0.076	-	1.6						
MTHFS	rs2586183	0.020	-	3.1									
SARDH	rs2073815										0.063	-	1.9
SHMT1	rs1979277	0.087	+	1.7									
TYMS	G>C in 3R allele VNTR										0.079	+	1.8
Remethylation													
BHMT	rs3733890										0.058	-	2
BHMT	rs672346										0.014	+	3.2
CUBN	rs1801231	0.039	+	2.1									

CUBN	rs1801239										0.064	-	1.9
CUBN	rs703075				0.051	+	2				0.040	-	2.2
MTHFR	rs1801133	0.002	-	5.2	0.026	+	2.8	0.006	+	4			
MTRR	rs1801394	0.011	-	3.5									
TCN2	rs1801198	0.059	+	2									
Methyl transfer													
AHCY	rs1205366							0.012	-	3.4			
DNMT1	rs8101626	0.003	+	4.5									
DNMT1	rs2290684	0.075	+	1.7									
DNMT3A	rs2276598				0.035	-	2.5	0.011	-	3.7			
ICMT	rs1802353										0.030	+	2.8
MAT2A	rs1078004										0.041	+	2.2
NNMT	rs1941404							0.083	+	1.6			
Transsulfuration													
CBS	844_845ins(68bp)				0.065	-	1.8	0.00050 [C]	-	6.3			
CBS	g.14037(31bp)16-21							0.018	-	2.9			
CTH	rs490574				0.086	+	1.6						
Other													
NOS1	4775(CA)14-22	0.105	-	1.7									

[A] indicates whether the minor allele is associated with an increase (+) or decrease (-) of metabolite concentration
[B] R^2 = explained variance
[C] statistically significant after correction for multiple testing

LD patterns

Examination of pair-wise LD patterns revealed that high LD (r^2>0.75) was present between rs682985 and rs526264 in *BHMT2*, and between rs2372536 and rs1997059, rs2372536 and rs4673991, and rs1997059 and rs4673991 in *ATIC*.

Serum folate

Here, we will focus on the ten DNA variants that showed the strongest association with each of the four metabolites as depicted in Table 3.2 (all other results available from corresponding author). The rs1801133 (*MTHFR*677C>T) was associated most strongly with serum folate (nominal p=0.002) and explained 5.2% of its variance. Other serum folate predictors included an intronic variant in *DNA (cytosine-5-)-methyltransferase 1* (*DNMT1*), for which also a second intronic variant was found among the strongest predictors (pair-wise r^2 between the SNPs = 0.706), and the non-synonymous rs1801394 (66A>G) in *5-methyltetrahydrofolate-homocysteine methyltransferase reductase* (*MTRR*). DNA variants located in genes encoding enzymes involved in uptake of folate and cobalamin were found among the strongest determinants of serum folate: rs1893007 in *FOLR1* and rs1801231 in *CUBN* and rs1801198 (776C>G) in *TCN2*. Also, DNA variants in genes encoding enzymes involved in the folate cycle were among the strongest predictors: the intronic rs2586183 in *5,10-methenyltetrahydrofolate synthetase* (*MTHFS*) and the non-synonymous rs1979277 (1420C>T) in *SHMT1*.

The complete results for the multi-locus set-based analysis are available as Supplementary Table 3.1. We found that for serum folate a combination containing rs1801133 in *MTHFR*, rs1801394 in *MTRR*, rs1801231 in *CUBN*, and rs1801198 in *TCN2*, all involved in the remethylation pathway, was most strongly associated; the sum-statistic almost reached statistical significance after correction for all performed tests (p=0.06597). There were no indications for association between (sums of) DNA variants involved in transsulfuration with serum folate.

Fasting and post-load plasma total homocysteine

The variance in fasting plasma tHcy explained by rs1801133 (*MTHFR*677C>T) was 2.8% (nominal p=0.026). We also found two variants of the transsulfuration pathway, the *CBS*844_845ins(68bp) and rs490574 in *CTH*, among the strongest determinants. Three DNA variants in *ATIC* in high LD (pair-wise r^2 >0.75) were among the strongest determinants of fasting plasma tHcy; the intronic rs4673991 explained 3.4% of the variance of plasma tHcy (nominal p=0.011).

For post-load plasma tHcy, both *CBS* variants measured in this study (pair-wise r^2 0.246) were represented among the strongest determinants; *CBS*844_845ins(68bp) explained 6.3% of the variation in residual post-load plasma tHcy and the association reached statistical significance after correction for multiple testing (p=0.044); *CBS* g.14037(31bp)16-21 explained 2.9% of variance. The rs1801133 (*MTHFR*677C>T)

showed stronger association with post-load than with fasting plasma tHcy, as did the synonymous rs2276598 in *DNA (cytosine-5-)methyltransferase 3 alpha* (*DNMT3A*). The intronic rs1205366 in *S-adenosylhomocysteine hydrolase* (*AHCY*) was nominally associated to post-load tHcy only. For both fasting and post-load plasma tHcy, rs4819205 in *formiminotransferase cyclodeaminase* (*FTCD*) was found among the strongest associated DNA variants. As can be seen in Table 3.2, the overlap in important genetic determinants for fasting/post-load plasma tHcy and serum folate was limited; only rs1801133 showed strong association with folate as well as tHcy.
For fasting and post-load plasma tHcy the multi-locus set-based analyses showed strongest associations for sums comprising DNA variants related to the transsulfuration pathway. For post-load tHcy, the sum containing only *CBS*844_845ins(68bp) showed statistical significant association after correction for all tests in all sets (p=0.032); for fasting plasma tHcy the sum including all DNA variants in *CBS* and *CTH* with exception of rs1021737 (also known as *CTH*1364G>T) was most strongly associated ($p_{\text{adjusted for tests within set}} = 0.111$).

Plasma methionine

Two uncorrelated non-synonymous DNA variants in *BHMT* were found among the strongest determinants for methionine concentrations, explaining 3.2 (nominal p=0.014) and 2.0% (nominal p=0.058) of variance. Also, two uncorrelated *CUBN* variants were identified; rs703075 (nominal p=0.040; also associated to fasting plasma tHcy) and rs1801239 (nominal p=0.064). Two DNA variants related to AdoMet-dependent methylation were also observed: rs1802353 in *isoprenylcysteine carboxyl methyltransferase* (*ICMT*) and rs1078004 in *methionine adenosyltransferase II, alpha* (*MAT2A*).

The strongest association in the set-based analysis was observed for the sum of the 'remethylation' set containing rs672346 in *BHMT*, rs703075 in *CUBN*, rs3733890 in *BHMT*, rs1801239 and rs1907362 in *CUBN*, but with high p-values ($p_{\text{adjusted for tests within set}} = 0.093$ and $p_{\text{adjusted for all tests}} = 0.380$).

Discussion

In the current study we aimed at identifying genetic determinants of three main intermediate metabolites of one-carbon metabolism that have been associated to disease: folate, homocysteine, and methionine. Our extensive one-carbon metabolism candidate-gene approach learned that multiple DNA variants from various pathways within the one-carbon metabolism each explained a small percentage of observed variance of metabolite concentrations in our population. New candidate DNA variants were identified among the strongest predictors of metabolite concentrations (e.g. variants in

CUBN, *DNMTs*, *FOLR1*, *FTCD*) in addition to candidates which have been studied by us and others previously (i.e. variants in *AHCY*, *BHMT*, *CBS*, *MTHFR*, *MTRR*, *SHMT1*, *TCN2*). We found that rs1801133 (*MTHFR*677C>T) was a main determinant for serum folate as well as, although to a lesser extent, fasting and post-load plasma tHcy. The latter was under strong influence of measured *CBS* variants. Overlap between DNA variants for fasting and post-load plasma tHcy were present, as expected given their correlation. The *MTHFR*677C>T was the only variant that showed very strong association to folate as well as tHcy.

Previous reports based on study samples including (a subset) of the 190 subjects that were used in the current study already revealed the association between *MTHFR*677TT and increased plasma fasting and post-load tHcy and decreased serum folate concentrations[40,41]. This is in accordance with published reports, the known reduced MTHFR activity associated with this variant, and the fact that *MTHFR*677C>T has been shown to result in accumulation of non-methyltetrahydrofolates at the expense of 5-methyltetrahydrofolate[21-26,49].

This study suggests that rs1801231 in *CUBN*, *TCN2*776C>G, and -8633C>T in *FOLR1*, all involved in remethylation of homocysteine and uptake and transport of dietary folate and cobalamin, are important contributors to variation in serum folate levels. Cubilin is, among others, involved in the uptake of cobalamin from the intestine. Cobalamin acts as cofactor for 5-methyltetrahydrofolate-homocysteine methyltransferase (MTR) and is needed for the 5-methyltetrahydrofolate-dependent remethylation of homocysteine to methionine that takes place in all cells with exception of erythrocytes. In addition, cobalamin serves as cofactor for methylmalonyl-CoA in the mitochondria[26,50]. *CUBN* variants also showed association to plasma tHcy and methionine concentrations; however, the specific DNA variants did not overlap with those strongly associated with folate concentrations. We recently reported on the association between a *CUBN* variant (rs1907362) and risk of spina bifida that showed nominal association to red blood cell folate and vitamin B_{12} concentrations but not to serum folate or plasma tHcy concentrations[31]. Others have not investigated DNA variants in *CUBN* in relation to concentrations of one-carbon metabolism intermediates to our knowledge. Transcobalamin (TC), encoded by TCN2, is required for cellular uptake of cobalamin from blood[26]. The common, non-synonymous rs1801198 (776C>G) in *TCN2* has been shown to decrease TC availability and has been associated to increased plasma tHcy, although inconsistently[25,26]. The increase in serum folate that we observed here has not been reported earlier; we did not find an association with plasma tHcy (nominal p=0.124 and 0.450 for fasting and post-load, respectively)[46]. The rare rs1893007, located in the promoter region of *FOLR1*, was associated with increased serum folate levels. This gene encodes the folate receptor α that is involved in cellular uptake of 5-methyl-THF; only few mutations in this gene have been reported[26,51], and this is the first study on the association between rs1893007 and folate concentrations.

The importance of the transsulfuration pathway in maintaining adequate homocysteine concentrations is evident; rare mutations in *CBS* are the main cause of the severely elevated plasma tHcy levels seen in homocystinuria[52]. We found an association between *CBS*844_845ins68 and decreased plasma tHcy concentrations, especially post-load, and a strong association between the g.14037(31bp)16-21 and post-load plasma tHcy concentrations in our current and previous[36,37] studies. Conflicting results for *CBS*844_845ins68 and plasma tHcy concentrations have been reported[25]; the association between g.14037(31bp)16-21 and post-load plasma tHcy concentrations has been confirmed in the Framingham Offspring Study[53] and our lab has shown variable CBS activity for different alleles of this polymorphism[36]. No associations with folate and methionine concentrations were found for these two *CBS* variants in the present study in agreement with a recent large sample size study[24]

Three correlated SNPs in *ATIC*, a gene involved in *de novo* purine synthesis, showed strong correlation with fasting plasma tHcy concentrations in the current study. Based on preliminary results of this study we have increased genotyping of the non-synonymous rs2372536 in *ATIC* in 241 additional Caucasian individuals in our original population-based sample. We have recently reported the findings of the association of this extended genotyping effort: we could not confirm the association to plasma tHcy[34] thereby raising doubt on the replicability of the SNP associations for this gene reported here.

Among the strongest determinants of methionine were rs672346 and rs3733890, two non-synonymous variants in *BHMT*. BHMT catalyzes an alternative remethylation route confined to the liver and kidneys in which betaine serves as a methyl donor[35]. The amino acid encoded by the codon in which the first SNP is located, is highly conserved. The second SNP has been previously investigated in a population including the current study subjects; no association with plasma tHcy was found[35], in accordance with other studies[25]. Fredriksen et al.[24] evaluated the association between rs3733890 and methionine concentrations but did not find a relation.

The results of our current study also suggested a role for genetic variation in methyl transfer-related genes and folate, homocysteine and methionine concentrations with little overlap in the implicated genes for the different metabolites: *DNMT1* variants contributed to variance in serum folate levels, variants in *DNMT3A* and *AHCY* explained over 3% variance each in post-load plasma tHcy, and a role for *ICMT* and *MAT2A* variants was indicated for methionine concentrations. A relation between one-carbon intermediate metabolites and impaired methylation potential has been described earlier[54]. For plasma tHcy, variation in nicotinamide N-methyltransferase (*NNMT*) has been suggested to influence plasma tHcy concentratons[55] and an association for the 324G>A variant in the catechol-O-methyltransferase gene (*COMT*) and plasma tHcy has been described by our group[56]. Also, a recent study by our group implicated

rs2295809 in *TRDMT1* in spina bifida etiology and showed indiciations for associations to increased folate concentrations, although not significantly[31].

Our data indicated that only few of our measured DNA variants may influence multiple measured metabolites (e.g. *MTHFR*677C>T and folate and homocysteine) while most show strong association to one metabolite only (e.g. variants in *CBS* and homocysteine). This is in line with findings from Fredriksen et al.[24] and has been supported by study of mouse models[57]. Indeed, the phenotypic correlations between the measured metabolites in the current study were consistent with influences of both shared genetic and/or environmental factors as well as metabolite-specific factors for folate and homocysteine. Pleiotropic effects of DNA variants may hamper the interpretation of Mendelian randomization studies. However, the existence of DNA variants that influence only one intermediate one-carbon metabolite of interest offers opportunities for studies based on Mendelian randomization to indicate which specific metabolite may have a role in disease causation[5].

We performed a candidate-gene association study for three metabolites with the largest number of measured candidate DNA variants until now. We preferably selected DNA variants that changed the function or regulation of the gene product, were potentially functional, and those that had already shown association with (any) disease or condition, thereby increasing the probability of actually genotyping causal variants. Genotyping of additional DNA variants in the selected genes increased our chance of indirectly finding genetic determinants via LD. However, we cannot assure that our study covered all relevant DNA variation in the selected genes. Ideally, future studies will optimize gene coverage, and, in addition, pathway coverage. Alternatively, genome-wide coverage may be aimed for to allow for identification of genetic determinants without reliance on current knowledge of potential candidate genes.

The complexity of the studied metabolic pathways urges the need for large sample sizes in genetic association studies of complex traits. Power calculations showed that under an additive genetic model we had 80% power to find a genetic variant that explains 4% and 3.2% of the variance in the trait at a nominal significance level of 0.05 and 0.10, respectively, which rendered our study underpowered for low-penetrance variants. Correction for multiple testing further diminishes power and increases the required number of subjects. The standardized measurements of fasting (and postload) levels of metabolites and adjustments for age and sex will have increased our power as the variation in metabolite concentration due to differences in age, sex, and, short-term dietary intake was reduced. However, future studies that include a larger study sample and replication of our findings are necessary. Larger samples also allow the analysis of (higher-order) interactions between DNA variants that will likely influence metabolite concentrations. In addition, gene-nutrient interactions and haplotype associations may be evaluated.

Here, we reported on the currently most extensive candidate-gene study for folate, homocysteine and methionine concentrations. Our study has established earlier findings and provided new leads for future research into genetic determinants of these one-carbon metabolites. We found little overlap in genetic determinants for folate, homocysteine, and methionine and mainly observed that DNA variants were strongly related to one metabolite, only.

Acknowledgments

Funding for this study was obtained from the Dutch Prinses Beatrix Fonds, grant MAR02-0206. This study was in part funded by Grant 2002B68 from the Netherlands Heart Foundation. Martin den Heijer, MD, PhD, is supported by a VENI grant from the Dutch Organization for Scientific Research (NW0).

References

1. Molloy AM, Scott JM. Folates and prevention of disease. Public Health Nutr. 2001;4(2B):601-609.
2. Keijzer MB, den Heijer M, Borm GF, Blom HJ, Vollset SE, Hermus AR, Ueland PM. Low fasting methionine concentration as a novel risk factor for recurrent venous thrombosis. Thromb Haemost. 2006;96:492-497.
3. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. Lancet. 1991;338:131-137.
4. Czeizel AE, Dudas I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation N Engl J Med. 1992;327:1832-1835.
5. Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32:1-22.
6. Daly LE, Kirke PN, Molloy A, Weir DG, Scott JM. Folate levels and neural tube defects. Implications for prevention. JAMA. 1995;274:1698-1702.
7. Homocysteine Studies Collaboration. Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA. 2002;288:2015-2022.
8. den Heijer M, Lewington S, Clarke R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost. 2005;3:292-299.
9. Bezemer ID, Doggen CJ, Vos HL, Rosendaal FR. No association between the common MTHFR 677C->T polymorphism and venous thrombosis: results from the MEGA study. Arch Intern Med. 2007;167:497-501.

10. Lewis SJ, Ebrahim S, Davey Smith G. Meta-analysis of MTHFR 677C->T polymorphism and coronary heart disease: does totality of evidence support causal role for homocysteine and preventive potential of folate? BMJ. 2005;331:1053-1059.
11. Bazzano LA, Reynolds K, Holder KN, He J. Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials. JAMA. 2006;296:2720-2726.
12. B-Vitamin Treatment Trialists' Collaboration. Homocysteine-lowering trials for prevention of cardiovascular events: a review of the design and power of the large randomized trials. Am Heart J. 2006;151:282-287.
13. Wang X, Qin X, Demirtas H, Li J, Mao G, Huo Y, Sun N, Liu L, Xu X. Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. Lancet. 2007;369:1876-1882.
14. Verhoef P, de Groot LC. Dietary determinants of plasma homocysteine concentrations. Semin Vasc Med. 2005;5:110-123.
15. Souto JC, Almasy L, Borrell M, Stone WH, Blanco-Vaca F, Soria JM, Blangero J, Fontcuberta J. Thromboplastin-thrombomodulin-mediated time and serum folate levels are genetically correlated with the risk of thromboembolic disease: results from the GAIT project. Thromb Heamost. 2002;87:68-73.
16. Cesari M, Burlina AB, Narkiewicz K, Sartori MT, Sacchetto A, Rossi GP. Are fasting plasma homocyst(e)ine levels heritable? A study of normotensive twins. J Investig Med. 2000;48:351-358.
17. Jee SH, Song KS, Shim WH, Kim HK, Suh I, Park JY, Won SY, Beaty TH. Major gene evidence after MTHFR-segregation analysis of serum homocysteine in families of patients undergoing coronary arteriography. Hum Genet. 2002;111:128-135.
18. den Heijer M, Graafsma S, Lee SY, van Landeghem B, Kluijtmans L, Verhoef P, Beaty TH, Blom H. Homocysteine levels--before and after methionine loading--in 51 Dutch families. Eur J Hum Genet. 2005;13:753-762.
19. Bathum L, Petersen I, Christiansen L, Konieczna A, Sørensen TI, Kyvik KO. Genetic and environmental influences on plasma homocysteine: results from a Danish twin study. Clin Chem. 2007;53:971-979.
20. Siva A, De Lange M, Clayton D, Monteith S, Spector T, Brown MJ. The heritability of plasma homocysteine, and the influence of genetic variation in the homocysteine methylation pathway. QJM. 2007;100:495-499.
21. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJH, den Heijer M, Kluijtmans LAJ, van den Heuvel LP, Rozen R. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet. 1995;10:111-113.
22. Brattström L, Wilcken DE, Ohrvik J, Brudin L. Common methylenetetrahydrofolate reductase gene mutation leads to hyperhomocysteinemia but not to vascular disease: the result of a meta-analysis. Circulation. 1998;98:2520-2526.

23. Molloy AM, Daly S, Mills JL, Kirke PN, Whitehead AS, Ramsbottom D, Conley MR, Weir DG, Scott JM. Thermolabile variant of 5,10-methylenetetrahydrofolate reductase associated with low red-cell folates: implications for folate intake recommendations. Lancet. 1997;349:1591-1593.
24. Fredriksen A, Meyer K, Ueland PM, Vollset SE, Grotmol T, Schneede J. Large-scale population-based metabolic phenotyping of thirteen genetic polymorphisms related to one-carbon metabolism. Hum Mutat. 2007;28:856-865.
25. Gellekink H, den Heijer M, Heil SG, Blom HJ. Genetic determinants of plasma total homocysteine. Semin Vasc Med. 2005;5:98-109.
26. van der Linden I, Afman LA, Heil SG, Blom HJ. Genetic variation in genes of folate metabolism and neural-tube defect risk. Proc Nut Soc. 2006;65:204-215.
27. den Heijer M, Blom HJ, Gerrits WB, Rosendaal FR, Haak HL, Wijermans PW, Bos GM. Is hyperhomocysteinaemia a risk factor for recurrent venous thrombosis? Lancet. 1995;345:882-885.
28. Fiskerstrand T, Refsum H, Kvalheim G, Ueland PM. Homocysteine and other thiols in plasma and urine: automated determination and sample stability. Clin Chem. 1993;39:263-271.
29. te Poele-Pothoff MT, van den Berg M, Franken DG, Boers GH, Jakobs C, de Kroon IF, Eskes TK, Trijbels JM, Blom HJ. Three different methods for the determination of total homocysteine in plasma. Ann Clin Biochem. 1995;32:218-220.
30. Holm PI, Ueland PM, Kvalheim G, Lien EA. Determination of choline, betaine, and dimethylglycine in plasma by a high-throughput method based on normal-phase chromatography-tandem mass spectrometry. Clin Chem. 2003;49:286-294.
31. Franke B, Vermeulen SH, Steegers-Theunissen RP, Coenen MJ, Schijvenaars MM, Scheffer H, den Heijer M, Blom HJ. An association study of 45 folate-related genes in spina bifida: involvement of Cubilin (CUBN) and tRNA aspartic acid methyltransferase 1 (TRDMT1). Birth Defects Res A Clin Mol Teratol. 2009;85:216-226.
32. N. Tõnisson, A. Kurg, K. Kaasik, E. Lohmussaar, A. Metspalu. Unravelling genetic data by arrayed primer extension. Clin Chem Lab Med. 2001;38:165-170.
33. den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet. 2001;109:121-124.
34. Gellekink H, Blom HJ, den Heijer M. Associations of common polymorphisms in the thymidylate synthase, reduced folate carrier and 5-aminoimidazole-4-carboxamide ribonucleotide transformylase/inosine monophosphate cyclohydrolase genes with folate and homocysteine levels and venous thrombosis risk. Clin Chem Lab Med. 2007;45:471-476.
35. Heil SG, Lievers KJ, Boers GH, Verhoef P, den Heijer M, Trijbels FJ, Blom HJ. Betaine-homocysteine methyltransferase (BHMT): genomic sequencing and relevance to hyperhomocysteinemia and vascular disease in humans. Mol Genet Metab. 2000;71:511-519.

36. Lievers KJ, Kluijtmans LA, Heil SG, Boers GH, Verhoef P, van Oppenraay-Emmerzaal D, den Heijer M. Trijbels FJ, Blom HJ. A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. Eur J Hum Genet. 2001;9:583-589.
37. Kluijtmans LA, Boers GH, Trijbels FJ, van Lith-Zanders HM, van den Heuvel LP, Blom HJ. A common 844INS68 insertion variant in the cystathionine beta-synthase gene. Biochem Mol Med. 1997;62:23-25.
38. Gellekink H, Blom HJ, van der Linden I, den Heijer M. Molecular genetic analysis of the human dihydrofolate reductase gene: relation with plasma total homocysteine, serum and red blood cell folate levels. Eur J Hum Genet. 2007;15:103-109.
39. Hol FA, van der Put NM, Geurds MP, Heil SG, Trijbels FJ, Hamel BC, Mariman ECM, Blom HJ. Molecular genetic analysis of the gene encoding the trifunctional enzyme MTHFD (methylenetetrahydrofolate-dehydrogenase, methenyltetrahydrofolate-cyclohydrolase, formyltetrahydrofolate synthetase) in patients with neural tube defects. Clin Genet. 1998;53:119-125.
40. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet. 1995;346:1070-1071.
41. Lievers KJ, Boers GH, Verhoef P, den Heijer M, Kluijtmans LA, van der Put NM, Trijbels FJ, Blom HJ. A second common variant in the methylenetetrahydrofolate reductase (MTHFR) gene and its relationship to MTHFR enzyme activity, homocysteine, and cardiovascular disease risk. J Mol Med. 2001;79:522-528.
42. Klerk M, Lievers KJ, Kluijtmans LA, Blom HJ, den Heijer M, Schouten EG, Kok FJ, Verhoef P. The 2756A>G variant in the gene encoding methionine synthase: its relation with plasma homocysteine levels and risk of coronary heart disease in a Dutch case-control study. Thromb Res. 2003;110:87-91.
43. van der Linden I, den Heijer M, Afman LA, Gellekink H, Vermeulen SH, Kluijtmans LA, Blom HJ. The methionine synthase reductase 66A>G polymorphism is a maternal risk factor for spina bifida. J Mol Med. 2006;84:1047-1054.
44. Heil SG, den Heijer M, Van Der Rijt-Pisa BJ, Kluijtmans LA, Blom HJ. The 894 G>T variant of endothelial nitric oxide synthase (eNOS) increases the risk of recurrent venous thrombosis through interaction with elevated homocysteine levels. J Thromb Haemost. 2004;2:750-753.
45. Heil SG, van der Put NM, Waas ET, den Heijer M, Trijbels FJ, Blom HJ. Is mutated serine hydroxymethyltransferase (SHMT) involved in the etiology of neural tube defects? Mol Genet Metab. 2001;73:164-172.
46. Lievers KJ, Afman LA, Kluijtmans LA, Boers GH, Verhoef P, den Heijer M, Trijbels FJ, Blom HJ. Polymorphisms in the transcobalamin gene: association with plasma homocysteine in healthy individuals and vascular disease patients. Clin Chem. 2002;48:1383-1389.

47. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263-265.
48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007;81:559-575.
49. Smulders YM, Smith DE, Kok RM, Teerlink T, Gellekink H, Vaes WH, Stehouwer CD, Jakobs C. Red blood cell folate vitamer distribution in healthy subjects is determined by the methylenetetrahydrofolate reductase C677T polymorphism and by the total folate status. J Nutr Biochem. 2007;18:693-699.
50. Christensen EI, Birn H. Megalin and cubilin: multifunctional endocytic receptors. Nat Rev Mol Cell Biol. 2002;3:256-266.
51. Börjel AK, Yngve A, Sjöström M, Nilsson TK. Novel mutations in the 5'-UTR of the FOLR1 gene. Clin Chem Lab Med. 2006;44:161-167.
52. Kraus JP, Janosík M, Kozich V, Mandell R, Shih V, Sperandeo MP, Sebastio G, de Franchis R, Andria G, Kluijtmans LA, Blom H, Boers GH, Gordon RB, Kamoun P, Tsai MY, Kruger WD, Koch HG, Ohura T, Gaustadnes M. Cystathionine beta-synthase mutations in homocystinuria. Hum Mutat. 1999;13:362-375.
53. Lievers KJ, Kluijtmans LA, Blom HJ, Wilson PW, Selhub J, Ordovas JM. Association of a 31 bp VNTR in the CBS gene with postload homocysteine concentrations in the Framingham Offspring Study. Eur J Hum Genet. 2006;14:1125-1129.
54. James SJ, Melnyk S, Pogribna M, Pogribny IP, Caudill MA. Elevation in S-adenosylhomocysteine and DNA hypomethylation: potential epigenetic mechanism for homocysteine-related pathology. J Nutr. 2002;132:2361S-2366S.
55. Souto JC, Blanco-Vaca F, Soria JM, Buil A, Almasy L, Ordoñez-Llanos J, Martín-Campos JM, Lathrop M, Stone W, Blangero J, Fontcuberta J. A genome-wide exploration suggests a new candidate gene at chromosome 11q23 as the major determinant of plasma homocysteine levels: results from the GAIT project. Am J Hum Genet. 2005;76:925-933.
56. Gellekink H, Muntjewerff JW, Vermeulen SH, Hermus AR, Blom HJ, den Heijer M. Catechol-O-methyltransferase genotype is associated with plasma total homocysteine levels and may increase venous thrombosis risk. Thromb Haemost. 2007;98:1226-1231.
57. Elmore CL, Matthews RG. The many flavours of hyperhomocysteinemia: insights from transgenic and inhibitor-based mouse models of disrupted one-carbon metabolism. Antioxid Redox Signal. 2007;9:1911-1921.

Supplementary Table 3.1 *Results of multi-locus set-based analysis for fasting serum folate, fasting plasma total homocysteine (tHcy), post-load plasma total homocysteine, and fasting methionine concentrations.*

Fasting serum folate						
Metabolic process	**Nr. of DNA variants in sum**	**Gene**	**DNA variant**	**p_0 [A]**	**p_1 [B]**	**p_2 [C]**
folate cycle	1	*MTHFS*	rs2586183	0.4393	0.5312	0.9845
folate cycle	2	*FOLR1*	rs1893007	0.3613	0.4583	0.96
folate cycle	3	*SHMT1*	rs1979277	0.3763	0.4773	0.9655
folate cycle	4	*MTHFD2*	rs1667627	0.3733	0.4748	0.9655
folate cycle	5	*GGH*	rs1031552	0.3523	0.4488	0.953
remethylation	1	*MTHFR*	rs1801133	0.02649	0.04498	0.2044
remethylation	2	*MTRR*	rs1801394	0.01299	0.02499	0.1104
remethylation	3	*CUBN*	rs1801231	0.01099	0.01949	0.09345
remethylation	4	*TCN2*	rs1801198	0.006997	0.01349	0.06597
remethylation	5	*CUBN*	rs2271461	0.007996	0.01599	0.07296
methyl transfer	1	*DNMT1*	rs8101626	0.05897	0.09245	0.3838
methyl transfer	2	*DNMT1*	rs2290684	0.1139	0.1644	0.5817
methyl transfer	3	*COQ3*	rs7755791	0.1464	0.2019	0.6697
methyl transfer	4	*ICMT*	rs1802353	0.1699	0.2359	0.7206
methyl transfer	5	*MGMT*	rs2020893	0.1804	0.2509	0.7471
transsulfuration	1	*CTH*	rs501939	0.9825	0.9885	1
transsulfuration	2	*CBS*	844_845ins(68bp)	0.973	0.9805	1
transsulfuration	3	*CTH*	rs3767205	0.9705	0.9785	1
transsulfuration	4	*CTH*	rs490574	0.968	0.9755	1
transsulfuration	5	*CTH*	rs1021737	0.9705	0.9785	1
other	1	*NOS1*	4775(CA)14-22	0.4473	0.5252	0.985
other	2	*NOS2A*	rs2297518	0.3608	0.4403	0.96
other	3	*NOS2A*	IVS1-2660(CCTTT)8-18	0.3163	0.3958	0.929
other	4	*NOS1*	IVS14+267(AAT)8-16	0.2554	0.3328	0.8706
other	5	*NOS3*	rs1799983	0.1954	0.2644	0.7801

[A] p_0: empirical p-value for average test-statistic.

[B] p_1: empirical p-value for average test-statistic corrected for all tests within this set.

[C] p_2: empirical p-value for average test-statistic corrected for all tests in all sets.

Fasting plasma tHcy						
Metabolic process	Nr. of DNA variants in sum	Gene	DNA variant	p_0 [A]	p_1 [B]	p_2 [C]
folate cycle	1	*ATIC*	rs4673991	0.2814	0.3713	0.8921
folate cycle	2	*FTCD*	rs4819205	0.3203	0.4143	0.924
folate cycle	3	*MTHFD1*	rs1950902	0.3398	0.4323	0.937
folate cycle	4	*GGH*	rs1031552	0.3928	0.4923	0.964
folate cycle	5	*GGH*	rs719235	0.4288	0.5247	0.9755
remethylation	1	*MTHFR*	rs1801133	0.2654	0.3488	0.8756
remethylation	2	*CUBN*	rs703075	0.2349	0.3148	0.8366
remethylation	3	*TCN2*	rs1801198	0.2459	0.3268	0.8491
remethylation	4	*CUBN*	rs1907362	0.2244	0.3018	0.8231
remethylation	5	*CUBN*	rs2271461	0.2034	0.2814	0.7926
methyl transfer	1	*DNMT3A*	rs2276598	0.3868	0.4928	0.961
methyl transfer	2	*AHCY*	rs1205366	0.5097	0.6122	0.99
methyl transfer	3	*MGMT*	rs2020893	0.5147	0.6157	0.9905
methyl transfer	4	*ICMT*	rs1802353	0.5042	0.6062	0.99
methyl transfer	5	*DNMT3A*	rs2289195	0.4823	0.5862	0.986
transsulfuration	1	*CBS*	844_845ins(68bp)	0.3423	0.4188	0.939
transsulfuration	2	*CTH*	rs490574	0.2174	0.2844	0.8116
transsulfuration	3	*CTH*	rs3767205	0.1294	0.1729	0.6342
transsulfuration	4	*CTH*	rs501939	0.09695	0.1379	0.5317
transsulfuration	5	*CBS*	g.14037(31bp)16-21	0.07546	0.1114	0.4558
other	1	*NOS1*	IVS14+267(AAT)8-16	0.5552	0.6282	0.992
other	2	*NOS2A*	rs2297518	0.4728	0.5487	0.9845
other	3	*NOS2A*	IVS1-2660(CCTTT)8-18	0.5257	0.5997	0.992
other	4	*NOS3*	rs1799983	0.5657	0.6397	0.993
other	5	*NOS3*	IVS4+245(GAAGTCTAGACC TGCTGCAGGGGTGAG)4-6	0.5822	0.6532	0.9945

[A] p_0: empirical p-value for average test-statistic.

[B] p_1: empirical p-value for average test-statistic corrected for all tests within this set.

[C] p_2: empirical p-value for average test-statistic corrected for all tests in all sets.

Post-load plasma tHcy						
Metabolic process	**Nr. of DNA variants in sum**	**Gene**	**DNA variant**	**p_0 [A]**	**p_1 [B]**	**p_2 [C]**
folate cycle	1	*FTCD*	rs4819205	0.3243	0.4238	0.9215
folate cycle	2	*ATIC*	rs4673991	0.3823	0.4818	0.9555
folate cycle	3	*FOLR1*	rs1893007	0.4173	0.5182	0.97
folate cycle	4	*ALDH1L1*	rs1127717	0.4573	0.5542	0.9795
folate cycle	5	*GGH*	rs1031552	0.4738	0.5737	0.9805
remethylation	1	*MTHFR*	rs1801133	0.08896	0.1339	0.5087
remethylation	2	*CUBN*	rs703075	0.1689	0.2404	0.7291
remethylation	3	*BHMT*	rs651852	0.1804	0.2539	0.7506
remethylation	4	*CUBN*	rs2271461	0.1719	0.2429	0.7351
remethylation	5	*BHMT2*	rs526264	0.1969	0.2744	0.7776
methyl transfer	1	*DNMT3A*	rs2276598	0.1179	0.1714	0.6032
methyl transfer	2	*AHCY*	rs1205366	0.05347	0.08546	0.3538
methyl transfer	3	*NNMT*	rs1941404	0.06247	0.09845	0.4003
methyl transfer	4	*AHCY*	rs819158	0.06147	0.09695	0.3958
methyl transfer	5	*ICMT*	rs1802353	0.06747	0.1064	0.4173
transsulfuration	1	*CBS*	844_845ins(68bp)	0.003498	0.006997	0.03198
transsulfuration	2	*CBS*	g.14037(31bp)16-21	0.003998	0.007996	0.03598
transsulfuration	3	*CTH*	rs3767205	0.004998	0.008996	0.04298
transsulfuration	4	*CTH*	rs490574	0.005997	0.01049	0.05147
transsulfuration	5	*CTH*	rs501939	0.007496	0.01249	0.06247
other	1	*NOS2A*	rs2297518	0.4543	0.5382	0.9785
other	2	*NOS3*	rs1799983	0.4463	0.5312	0.977
other	3	*NOS2A*	IVS1-2660(CCTTT)8-18	0.4498	0.5347	0.9775
other	4	*NOS1*	4775(CA)14-22	0.4118	0.4993	0.968
other	5	*NOS1*	IVS14+267(AAT)8-16	0.3973	0.4818	0.9625

[A] p_0: empirical p-value for average test-statistic.

[B] p_1: empirical p-value for average test-statistic corrected for all tests within this set.

[C] p_2: empirical p-value for average test-statistic corrected for all tests in all sets.

Fasting plasma methionine						
Metabolic process	**Nr. of DNA variants in sum**	**Gene**	**DNA variant**	**p_0 [A]**	**p_1 [B]**	**p_2 [C]**
folate cycle	1	*GART*	rs8971	0.7341	0.8036	1
folate cycle	2	*SARDH*	rs2073815	0.6887	0.7706	1
folate cycle	3	*ALDH1L1*	rs1127717	0.6097	0.6987	0.9975
folate cycle	4	*TYMS*	CCGCGCCACTTGGCCTGCCTCCGTC CCG>CCGCGCCACTTCGCCTGCCTC CGTCCCG	0.5402	0.6457	0.9925
folate cycle	5	*SARDH*	rs2073817	0.4883	0.5937	0.986
remethylation	1	*BHMT*	rs672346	0.1919	0.2619	0.7756
remethylation	2	*CUBN*	rs703075	0.1644	0.2329	0.7191
remethylation	3	*BHMT*	rs3733890	0.1244	0.1839	0.6267
remethylation	4	*CUBN*	rs2271461	0.07896	0.1249	0.4623
remethylation	5	*CUBN*	rs1907362	0.05947	0.09345	0.3798
methyl transfer	1	*ICMT*	rs1802353	0.3073	0.3978	0.9065
methyl transfer	2	*MAT2A*	rs1078004	0.2334	0.3123	0.8376
methyl transfer	3	*DNMT3A*	rs2276598	0.2249	0.3028	0.8241
methyl transfer	4	*AHCY*	rs1205366	0.2214	0.2989	0.8216
methyl transfer	5	*TRDMT1*	rs2295809	0.2129	0.2884	0.8091
transsulfuration	1	*CTH*	rs3767205	0.6977	0.7596	1
transsulfuration	2	*CTH*	rs1021737	0.6807	0.7426	1
transsulfuration	3	*CTH*	rs490574	0.6902	0.7506	1
transsulfuration	4	*CTH*	rs501939	0.7051	0.7651	1
transsulfuration	5	*CBS*	844_845ins(68bp)	0.7226	0.7801	1
other	1	*NOS1*	4775(CA)14-22	0.5262	0.6082	0.9915
other	2	*NOS2A*	rs2297518	0.5352	0.6177	0.9925
other	3	*NOS3*	IVS4+245(GAAGTCTAGACCTGCTG CAGGGGTGAG)4-6	0.5472	0.6292	0.9935
other	4	*NOS1*	IVS14+267(AAT)8-16	0.5802	0.6617	0.9965
other	5	*NOS3*	rs1799983	0.5787	0.6592	0.9965

[A] p_0: empirical p-value for average test-statistic.

[B] p_1: empirical p-value for average test-statistic corrected for all tests within this set.

[C] p_2: empirical p-value for average test-statistic corrected for all tests in all sets.

Sita HHM Vermeulen
Sandra G Heil
Henkjan Gellekink
Ivon JM van der Linden
Ad RMM Hermus
Leo AJ Kluijtmans
Henk J Blom
Martin den Heijer

Submitted

CHAPTER 4

Multi-locus analysis of candidate DNA variants for plasma homocysteine concentration: identification of highly associated multi-locus genotype

Abstract

The application of multi-locus analysis techniques may aid in the elucidation of the genetic aetiology of multi-factorial traits. Elevated plasma total homocysteine (tHcy) concentration is a multi-factorial risk factor for cardiovascular diseases, venous thrombosis, congenital defects, and neurodegenerative diseases. Its genetic background is unresolved. Our study objective was to evaluate candidate DNA variants for plasma tHcy concentration using multi-locus analysis. Thirty-six DNA variants in 19 genes related to homocysteine metabolism were measured in a study population comprising 461 Caucasian individuals. Single locus, haplotype, and multi-locus genotype analysis using logic regression were performed for fasting and post-methionine load plasma tHcy concentration. Single locus results were previously reported and suggested main roles for *MTHFR*677C>T, *CBS*844ins68, *CBS* 31bp VNTR, and *COMT*324G>A. No haplotype effects independent of single locus effects were found. Logic regression identified a multi-locus genotype {*CBS* 31bp VNTR 18-18 and *FOLH1*1561CC and (*MTHFD*2011GG or *BHMT*595GA) and *MTHFR*677CT/TT} that is present in 10% of the study population and associated with 5 µmol/L higher fasting plasma tHcy. It explained ~17% of trait variance in our population; the 5 DNA variants separately explained ~5% of plasma tHcy variance. *MTHFR*677C>T and *CBS* 31bp VNTR are main genetic determinants of plasma tHcy concentration with modest marginal effects. A deleterious combination of genotypes at five measured DNA variants predisposed to highly elevated plasma tHcy concentration. Results of this study demonstrate the additional value of multi-locus analysis in the elucidation of the genetic aetiology of plasma tHcy concentration.

Introduction

An increased awareness of potential advantages of a joint evaluation of multiple loci for multi-factorial traits has arisen in the last years and sophisticated techniques for analyzing multi-locus effects have been developed[1]. Indeed, the analysis of single DNA variants only allows identification and characterization of those variants that show an independent, marginal contribution to the phenotype of interest, either directly or indirectly via linkage disequilibrium (LD) to another, causal DNA variant. However, interactions between DNA variants, that show small or no marginal individual effects, are expected to contribute to the phenotypic variance in multi-factorial traits. These may concern interactions between multiple DNA variants located in the same genes or separate genes, or specific interactions between alleles of variants that are located on the same chromosomal background (i.e. in *cis*). The latter type of interaction can be explicitly identified using haplotype analysis, in which the combination of alleles on multiple (linked) variants on a chromosome is evaluated for association to the trait. Also, analysis of haplotypes can be advantageous compared to single locus analysis when unmeasured, causal genetic variants show stronger LD to the haplotype than to the single genetic variants themselves[2].

Elevated plasma total homocysteine (tHcy) concentrations are associated with increased risk for several multi-factorial disorders including cardiovascular diseases, venous thrombosis, congenital defects, and neurodegenerative diseases[3-8]. However, recent secondary intervention trials with plasma tHcy-lowering therapy did not show conclusive results for mortality and vascular disease risk and raised doubt on the causality of these associations[9-11]. Clarification of the genetic aetiology of plasma tHcy can help to elucidate the pathophysiological pathways underlying the disease associations and lead to an increased understanding of disease aetiology. Plasma tHcy is a multi-factorial trait itself; the inter-individual variation in plasma tHcy in the general population is caused by environmental and genetic determinants that interact with one another[12,13].

In the past years we and others have been involved in deciphering the genetic aetiology of plasma tHcy through a candidate gene approach in which variants of genes encoding key enzymes in homocysteine metabolism were evaluated for their association with plasma tHcy[14]. The best established genetic determinant of plasma tHcy concentration in the general population is the 677C>T single nucleotide polymorphism (SNP) in the methylenetetrahydrofolate reductase (*MTHFR*) gene that plays a regulatory role in the remethylation of homocysteine to methionine[15]. Other, common and uncommon, candidate gene variants have been studied. However, contradictive association findings exist and the genetic background of plasma tHcy remains unresolved[14]. Generally, genetic association studies for plasma tHcy concentrations examined one or few deoxyribonucleic acid (DNA) variants in one or few genes and

entailed statistical analyses in which each measured variant was analysed separately. The aim of the present study was to examine the DNA variants that we have measured over the last years in our study population for association with plasma tHcy using a *multi-locus* analysis approach and perform haplotype and logic regression analysis. For completeness of the results, we also present single locus associations.

Our candidate gene approach entailed the study of 36 DNA variants in 19 candidate genes in homocysteine metabolism (Figure 4.1) in a population comprising 500 subjects in which fasting as well as post-methionine load (hereafter named post-load) plasma tHcy have been measured. These two homocysteine phenotypes have a different genetic aetiology[13]. The associations between the single measured DNA variants and plasma tHcy have been described earlier by our group[16-36]. In few cases, associations with composite genotypes within a gene[21,30-32], between genes[16,19], or haplotype associations[20] were evaluated. Main findings of our studies include the confirmation of the association between *MTHFR*677C>T and plasma tHcy[31], and the identification of the 1444_1467+7(16_21) (known as 31 base-pair (bp) VNTR) in the gene encoding cystathionine ß-synthase (*CBS*) (30) and the 324G>A SNP in the catechol-O-methyltransferase gene (*COMT*)[20] as genetic determinants of plasma tHcy. In addition, we found associations between fasting and/or post-load plasma tHcy and 80G>A in solute carrier family 19 (folate transporter), member 1 (*SLC19A1=RFC1*)[19], the 146_164 deletion in dihydrofolate reductase (*DHFR*)[18], and the 45 bp insertion/deletion variant (1032-1077dup) in uncoupling protein 2 (*UCP2*)[25]. Also, interactions between 346C>G in 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase (*ATIC*) and *SLC19A1*80G>A[19], and between the 31bp VNTR in *CBS* and *MTHFR*677C>T were described[16].

The current study will focus on multi-locus analyses with the aim to increase insight in the genetic aetiology of plasma tHcy and the role of interplay between gene variants, within and among genes. Moreover, it will give a quantitative overview of the results of our candidate gene approach.

Materials and methods

Study population

The study population has been described extensively before [37]. In short, 500 healthy individuals were recruited in 1993 via a general practice in The Hague. Thirty-nine non-Caucasian subjects were excluded for this study. Of the remaining 461 subjects 42% was male and median age was 50 years (range 21 – 84 years). Median (5^{th} – 95^{th} percentile) fasting (n=461) and post-load plasma tHcy levels (n=457) were 10.7 (5.8 - 18.4) µmol/L and 37.7 (23.6 – 66.2) µmol/L, respectively.

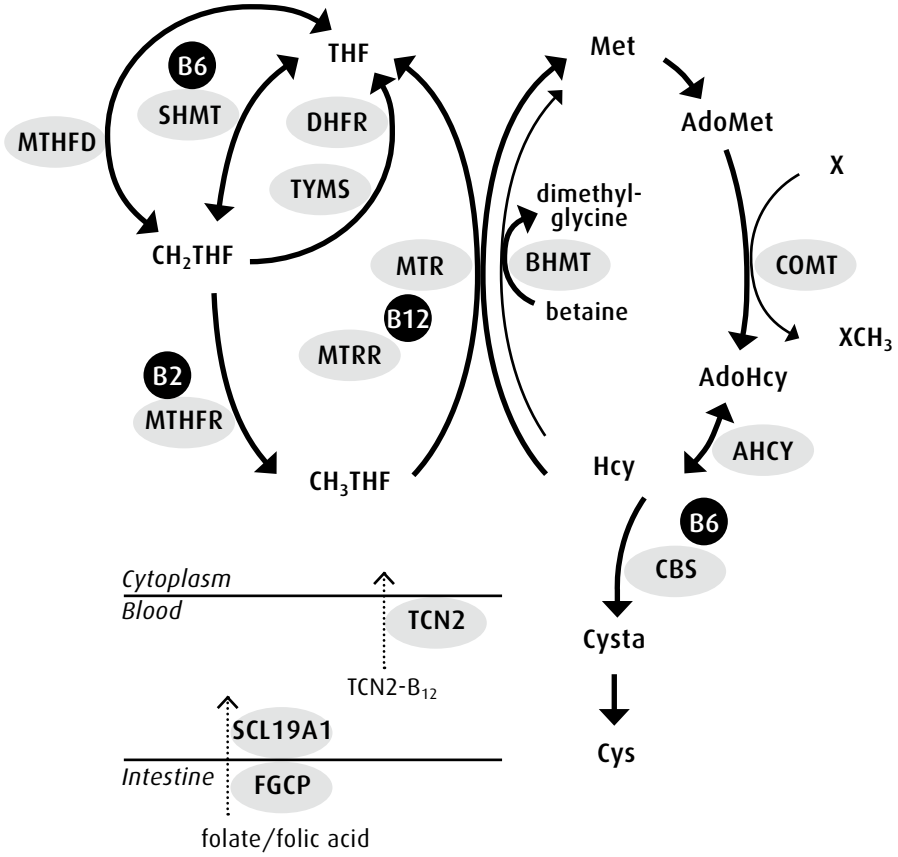

Figure 4.1 *Simplified representation of homocysteine metabolism. Enzymes and receptors for which DNA variants in the encoding genes are measured in the current study are depicted (with exception of ATIC, GCLM, NOS3, and UCP2).*

AdoHcy: S-adenosylhomocysteine; AdoMet: S-adenosylmethionine; AHCY: S-adenosylhomocysteine hydrolase; ATIC: 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase; B2: vitamin B_2 (riboflavin); B6: vitamin B_6 (pyridoxine); B12: vitamin B_{12} (cobalamin); BHMT: betaine-homocysteine S-methyltransferase; CBS: cystathionine-beta-synthase; CH_2THF: 5,10-methylenetetrahydrofolate; CH_3THF: 5-methyltetrahydrofolate; COMT: catechol-O-methyltransferase; Cys: cysteine; Cysta: cystathionine; DHFR: dihydrofolate reductase; FGCP: folylpoly-glutamyl carboxypeptidase; GCLM: glutamate-cysteine ligase, modifier subunit; Hcy: homocysteine; Met: methionine; MTHFD: methylenetetrahydrofolate dehydrogenase; MTHFR: methylenetetrahydrofolate reductase; MTR: methionine synthase; MTRR: methionine synthase reductase; NOS3: nitric oxide synthase 3; SCL19A1: solute carrier family 19 (folate transporter), member 1; SHMT: serine hydroxymethyltransferase; TCN2; transcobalamin II; THF: tetrahydrofolate; TYMS: thymidylate synthetase; UCP2: uncoupling protein 2; X: methylacceptor.

Table 4.1 *Characteristics of DNA variants analysed in this study.*

Gene symbol	Gene name	Locus	Accession no.	SNP[1] ID	DNA[2] variant[3]	Amino acid change	Reference[4]	No. genotyped	Minor allele	MAF[5]	HWE[6] P
AHCY	S-adenosylhomocysteine hydrolase	20cen-q13.1	NM_000687	-	-34C>T	-	17	392	T	0.018	0.719
AHCY	S-adenosylhomocysteine hydrolase	20cen-q13.1	NM_000687	rs13043752	112C>T	Arg38Trp	17	404	T	0.029	0.226
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/ IMP cyclohydrolase	2q35	NM_004044	rs2372536	346C>G	Thr116Ser	19	381	C	0.332	0.644
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	NM_001713	-	595G>A	Gly199Ser	21	407	A	0.006	0.901
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	NM_001713	rs3733890	716G>A	Arg239Gln	21	393	A	0.305	0.696
BHMT	betaine-homocysteine methyltransferase	5q13.1-q15	NM_001713	-	1218G>T	Glu406His	21	374	T	0.001	0.980
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	-	-5697(GT)10-20	-	34	393	-	-	0.599
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	rs234706	699C>T	-	34	378	T	0.365	0.561
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	rs5742905	833T>C	Ile278Thr	28	384	C	0.005	0.918
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	-	844_845ins(68bp)[7]	-	29	384	ins	0.091	0.265
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	rs1801181	1080C>T	-	34	416	T	0.365	0.115
CBS	cystathionine-beta-synthase	21q22.3	NM_000071	-	1444_1467+7(16_21) (31bp VNTR)[8]	-	30	387	-	-	0.022
COMT	catechol-O-methyltransferase	22q11.21	NM_007310	rs2097603	-278A>G	-	20	390	G	0.488	0.992
COMT	catechol-O-methyltransferase	22q11.21	NM_007310	rs4633	36T>C	-	20	383	C	0.456	0.758
COMT	catechol-O-methyltransferase	22q11.21	NM_007310	rs4680	324G>A	Val108Met	20	393	G	0.472	0.770
COMT	catechol-O-methyltransferase	22q11.21	NM_007310	rs174699	g.25150T>C	-	20	390	C	0.065	0.783
DHFR	dihydrofolate reductase	5q11.2-q13.2	NM_000791	-	-388(9bp)3-9[9]	-	18	385	-	-	0.128
DHFR	dihydrofolate reductase	5q11.2-q13.2	NM_000791	-	146_164del	-	18	301	del	0.427	0.329
FOLH1	glutamate carboxypeptidase II	11p11.2	NM_001014986	-	1561C>T	His475Tyr	33	412	T	0.057	0.129
GCLM	glutamate-cysteine ligase, modifier subunit	1p22.1	NM_002061	-	-588C>T	-	24	390	T	0.227	0.400

MTHFD	methylenetetrahydrofolate dehydrogenase	14q24	NM_005956	rs2236225	2011G>A	Arg653Gln	26	386	A	0.451	0.186
MTHFR	5,10-methylenetetrahydrofolate reductase	1p36.3	NM_005957	rs1801133	665C>T[10]	Ala222Val	36 (see also 31)	445	T	0.296	0.346
MTHFR	5,10-methylenetetrahydrofolate reductase	1p36.3	NM_005957	rs1801131	1286A>C[10]	Glu429Ala	31	438	C	0.334	0.391
MTR	5-methyltetrahydrofolate-homocysteine methyltransferase	1q43	NM_000254	rs1805087	2756A>G	Asp919Gly	27	408	G	0.140	0.419
MTRR	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	5p15.31	NM_002454	rs1801394	66A>G	Ile22Met	35	413	A	0.427	0.909
NOS3	nitric oxide synthase 3	7q36	NM_000603	rs1799983	894G>T	Asp298Glu	23	403	T	0.305	0.899
SHMT1	serine hydroxymethyltransferase 1	17p11.2	NM_004169	rs1979277	1420C>T	Leu474Phe	22	373	T	0.306	0.488
SHMT2	serine hydroxymethyltransferase 2	12q13.2	NM_005412	-	1721_1722insTCTT	-	22	418	ins	0.028	0.211
SLC19A1	solute carrier family 19 (folate transporter), member 1	21q22.3	NM_194255	rs1051266	80G>A	His27Arg	19	398	A	0.379	0.484
TCN2	transcobalamin II	22q12.2	NM_000355	rs11557600	280G>A	Gly94Ser	32	380	A	0.009	0.856
TCN2	transcobalamin II	22q12.2	NM_000355	rs1801198	776C>G	Arg259Pro	32	399	G	0.456	0.311
TCN2	transcobalamin II	22q12.2	NM_000355	rs9621049	1043C>T	Ser348Phe	32	403	T	0.115	0.248
TCN2	transcobalamin II	22q12.2	NM_000355	rs4820889	1196G>A	Arg399Gln	32	405	A	0.019	0.704
TYMS	thymidylate synthetase	18p11.32	NM_001071	-	-220(28bp)2-4[11]	-	19	389	-	-	0.546
TYMS	thymidylate synthetase	18p11.32	NM_001071	-	1494_1499delTTAAAG	-	19	398	del	0.351	0.808
UCP2	uncoupling protein 2	11q13	NM_003355	-	1032-1077dup	-	25	355	ins	0.276	0.189

[1] Single nucleotide polymorphism
[2] Deoxyribonucleic acid
[3] Position at cDNA according to nomenclature of den Dunnen and Antonarakis[39] unless stated otherwise
[4] Reference to paper in which association between DNA variant and plasma tHcy for (part of) this population was first reported
[5] Minor allele frequency; given for di-allelic DNA variants only
[6] Hardy-Weinberg Equilibrium
[7] 68 bp: CATCCAGGTGGGGTTTTGCTGGGCTTGAGCCCTGAAGCCGCGCCCTCTGCAGATCATTGGGGTGGAT
[8] Heterogeneous repeat units are not presented here; see Lievers et al.[30]
[9] Heterogeneous 9 bp repeat units are not presented here; see Gellekink et al.[18]
[10] Sequence variants will further be referred to as *MTHFR*677C>T and *MTHFR*1298A>C as is consistently done in other publications
[11] 28 bp: CCGCGCCACTTGGCCTGCCTCCGTCCCG

Genotyping and phenotypic measurements

Blood samples were drawn from the antecubital vein and DNA was extracted using standard procedures[38]. We studied 36 DNA variants in 19 candidate genes that were selected for their essential roles in homocysteine and folate metabolism. Selection of the DNA variants within candidate genes was mainly based on gene sequencing in patient populations or on other published findings with priority for variants with (potential) functionality. Genotyping was mostly done by means of Polymerase Chain Reaction (PCR) and Restriction Fragment Length Polymorphism (RFLP) analysis techniques over the last 13 years at different time points and resulted in different numbers of successfully genotyped individuals for the DNA variants. Details regarding the DNA variants, their selection, and genotyping procedures can be found in the papers by our group in which they were first described (Table 4.1).

Venous blood samples were collected after an overnight fast. Total fasting plasma tHcy was measured by automated high-pressure liquid chromatography (HPLC) with reverse phase separation and fluorescent detection[40]. A second plasma tHcy assessment took place 6 hours after an oral methionine loading (100 mg L-methionine per kg bodyweight dissolved in 200 ml orange juice). During these 6 hours, the subjects obtained a protein-restricted diet. Serum creatinine was measured with a Kodak Ektachem Processor.

Statistical genetic analyses

Single locus analysis Deviations from Hardy-Weinberg equilibrium (HWE) for di-allelic and multi-allelic variants were evaluated using Chi-square tests and the Monte Carlo method[41], respectively. The distribution of plasma tHcy was skewed to the right and tHcy values were natural logarithmically transformed prior to all analyses. Genotypic differences in fasting and post-load plasma tHcy were tested using multivariable linear regression techniques under assumption of a genotypic and additive gene model. Regression analysis was performed for unadjusted plasma tHcy and for plasma tHcy adjusted for age, sex and creatinine levels; these latter variables are known determinants of plasma tHcy. Empirical nominal *P*-values were based on a permutation procedure using 1000 replications. We used Stata (version 9.0) and the GENHW add-on package[42] to perform these analyses. All reported *P*-values are two-sided; we considered nominal *P*-values <0.05 to indicate statistically significant results.

Haplotype analysis LD coefficients between gene variants located within one gene were estimated and haplotype block structure, based on the algorithm described by Gabriel et al[43], was mapped using Haploview[44] (http://www.broad.mit.edu/personal/jcbarret/haploview/index.php). Haplotype analysis for unadjusted and adjusted plasma tHcy was performed using the Whap program (http://pngu.mgh.harvard.edu/~purcell/whap/)[45]. We first applied an omnibus test, jointly testing all haplotypes, and then evaluated haplotype-specific effects relative to a reference haplotype.

Haplotypes with a frequency of less than 2% were excluded. The multi-allelic variants were recoded into multiple di-allelic variants prior to LD and haplotype analysis. We used permutation testing procedures to generate empirical nominal *P*-values using 1000 replications. Again, all reported *P*-values are two-sided and *P*-values <0.05 were considered statistically significant.

Logic regression Logic regression is an explorative regression method designed to identify a Boolean (logical) combination (or combinations) of original binary variables associated with the outcome of interest. The method has been extensively described before[46,47]. Our goal was to find combinations of DNA variants that may interact with one another and that are associated with plasma tHcy. We used the linear logic regression technique as implemented as an R package (http://bear.fhcrc.org/~ingor/logic/) (R version 2.3.1). DNA variants were recoded into binary variables using a recessive and dominant coding. Indicator variables were generated for the 31 bp VNTR and the GT simple tandem repeat (STR) (-5697(GT)10-20) in *CBS*, and for the 9 bp repeat (-388(9bp)3-9) in *DHFR* with omission of low frequency genotypes. Model selection was guided by ten-fold cross-validation (as integrated in the logic regression package) for model sizes ranging from one to three trees with a maximum of eight leaves in total. Logic regression was limited to subjects with complete data for *all* included DNA variants. To increase the number of analysed subjects, we selected only those DNA variants that showed association with fasting or post-load plasma tHcy with *P*-value < 0.10 and >1% variety for the recoded binary variable. We repeated single locus analysis in the logic regression sub-sample to allow for fair comparison between the single locus and multi-locus genotype analysis. We explored marginal and interactive contributions to variation in plasma tHcy of the DNA variants identified in the logic regression using analysis of variance (ANOVA) models.

Results

Table 4.1 shows characteristics of the analysed candidate genes and DNA variants. We measured four repeat variants, five insertion/deletion variants, and 27 single nucleotide di-allelic variants, of which 4 had a minor allele frequency (MAF) <1%. Only the 31 bp VNTR in *CBS* showed slight deviation from HWE ($P = 0.022$). Twenty DNA variants resulted in an amino acid change and most of the variants are known or suspected to affect function of the encoded protein (see our earlier papers). The number of successfully genotyped individuals varied from 301 to 445.

Table 4.2 *Genotypic single locus association results for unadjusted plasma tHcy concentration for DNA variants with P <0.05.*

Gene symbol	DNA variant	Genotype	N (%)	Fasting tHcy	Post-load tHcy
				% difference (95% CI)	% difference (95% CI)
BHMT	595G>A	GG	402 (98.8)	- [1]	- [1]
		GA	5 (1.2)	47.7 (5.7 to 106.3) 2	11.3 (-16.0 to 47.7)
		AA	- (-)	-	-
CBS	699C>T	CC	155 (41.0)	-	
		CT	170 (45.0)	-9.4 (-16.2 to -1.9) [2]	-6.9 (-12.8 to -0.6) [2]
		TT	53 (14.0)	-5.6 (-15.7 to 5.7)	-6.8 (-15.1 to 2.3)
CBS	833T>C	TT	380 (99.0)	-	-
		TC	4 (1.0)	42.9 (-2.1 to 108.6) [3]	106.6 (51. 9 to 181.0)
		CC	0 (0)	-	-
CBS	844_845ins68	del del	319 (83.1)	-	-
		del ins	60 (15.6)	-10.8 (-19.7 to -0.9)	-10.9 (-18.5 to -2.7)
		ins ins	5 (1.3)	-32.8 (-52.0 to -5.8)	-18.1 (-38.2 to 8.5)
CBS	31 bp VNTR[4]	18 18	233 (60.2)	-	-
		17 18	52 (13.4)	-11.5 (-21.0 to -1.0)	-11.7 (-19.6 to -3.0)
		18 19	62 (16.2)	-8.4 (-17.5 to 1.7)	-9.5 (-17.0 to -1.3)
COMT	324G>A	AA	111 (28.2)	-	-
		AG	193 (49.1)	-5.1 (-13.2 to 3.7)	-1.5 (-8.5 to 6.1)
		GG	89 (22.6)	-11.7 (-20.6 to -1.8)	-9.3 (-17.0 to -0.9)
DHFR	146_164del	ins ins	103 (34.2)	-	-
		ins del	139 (46.2)	-1.0 (-9.6 to 8.4)	4.7 (-3.3 to 13.4)
		del del	59 (19.6)	-14.4 (-23.6 to -4.1)	-3.3 (-12.6 to 7.0)
FOLH1	1561C>T	CC	368 (89.3)	-	-
		CT	41 (10.0)	-4.3 (-15.2 to 8.1)	3.5 (-6.7 to 14.8)
		TT	3 (0.7)	-42.8 (-62.7 to -12.3)	-24.4 (-47.5 to 8.7)
GCLM	-588C>T	CC	236 (60.5)	-	-
		CT	131 (33.6)	-9.0 (-16.2 to -1.3)	-1.0 (-7.6 to 6.1)
		TT	23 (5.9)	-13.8 (-26.8 to 1.6)	-8.5 (-20.5 to 5.3)
MTHFR	677C>T	CC	225 (50.6)	-	-
		CT	177 (39.8)	3.8 (-3.6 to 11.6)	7.4 (1.0 to 14.2)
		TT	43 (9.7)	20.1 (6.4 to 35.6)	16.0 (4.45 to 28.7)
SLC19A1	80G>A	GG	150 (37.7)	-	-
		GA	194 (48.7)	-6.0 (-13.3 to 1.9)	-9.0 (-14.9 to -2.6)
		AA	54 (13.6)	-9.9 (-19.9 to 1.4)	1.1 (-8.5 to 11.7)

[1] Reference category; [2] P-value >0.05 after adjustment for age, sex and creatinine; [3] P-value <0.05 after adjustment for age, sex and creatinine; [4] Only the three most frequent genotype groups are presented

Single locus association

Here, we only discuss the DNA variants that were associated to fasting or post-load plasma tHcy with nominal *P*-value < 0.05 (Table 4.2) (the results of all single locus analyses can be requested from the corresponding author). Table 4.2 shows that four *CBS* variants were found among the strongest association results. Relative difference in plasma tHcy for the *MTHFR*677TT genotype compared to *MTHFR*677CC genotype was 20% (crude geometric mean tHcy levels were 12.3 µmol/L and 10.2 µmol/L, respectively); other common DNA variants showed smaller effects. The low-frequency variants *BHMT*595G>A, *CBS*833T>C, *CBS*844_845ins68 and *FOLH1*1561C>T did show large effect on plasma tHcy in our population. The minor alleles of the DNA variants in Table 4.2 were all associated with decreased plasma tHcy, except for the *MTHFR*677 T allele, the *CBS*833 C allele and increasing length of the 31 bp VNTR in *CBS*. Results for fasting and post-load plasma tHcy differed slightly. For instance, the decreases in plasma tHcy of *GCLM*-588C>T, *DHFR*146_164del, and *FOLH1*1561C>T, were less strong for post-load plasma tHcy. The 833T>C variant and the 31 bp VNTR in *CBS*, however, were associated with stronger increases in post-load compared to fasting levels.

Haplotype association

*MTHFR*677C>T and *MTHFR*1298A>C were located in one haplotype block, as were the 844_845ins68, the 1080C>T, and the 31 bp VNTR in *CBS*, and the 36T>C and 324A>G variant in *COMT* which indicated that alleles of these variants are inherited together on the same chromosomal background. Strong LD in terms of r^2 was only found for the latter two variants in *COMT* (r^2=0.892) and the '17' allele of the 31 bp VNTR and the 844_845ins68 variant in *CBS* (r^2=0.833) which indicated that these variants 'tag' each other well and one can serve as a good proxy for the other (data not shown).

Statistically significant haplotype block effects for *MTHFR*, *CBS* and *COMT* for fasting plasma tHcy were identified (Table 4.3). Conditional analysis showed that haplotype effects were no longer significant after conditioning on *MTHFR*677C>T, 844_845ins68 or 31 bp VNTR in *CBS*, and *COMT*324G>A, respectively. This indicated that haplotype associations did not contribute to plasma tHcy in addition to the marginal single variant effects. The same was observed for post-load plasma tHcy; after adjustment for the *MTHFR*677C>T and the *CBS* 31 bp VNTR the *P*-values of the haplotype tests increased to non-significant values. Haplotype analyses including *all* measured loci within a gene, irrespective of block structure, showed the same patterns (data not shown).

Table 4.3 *Haplotype association results for haplotype blocks in CBS, COMT, and MTHFR and plasma tHcy concentration. Number of informative individuals for the haplotype analysis were 461, 429 and 455 for MTHFR, CBS and COMT, respectively.*

Gene	Variants				Fasting tHcy		Fasting tHcy adjusted for age, sex, creatinine		Post-load tHcy		Post-load tHcy adjusted for age, sex, creatinine	
	Haplotype			Frequency	P omnibus	Difference (%) (95% CI)	P omnibus	Difference (%) (95% CI)	P omnibus	Difference (%) (95% CI)	P omnibus	Difference (%) (95% CI)
CBS	844	1080	VNTR									
	del	C	18	0.427	0.018	-	0.015	-	0.003	-	0.001	-
	del	T	18	0.371		-3.4 (-8.6 to 2.1)		-3.5 (-8.2 to 1.5)		-3.8 (-8.3 to 1.0)		-3.8 (-8.3 to 0.9)
	del	C	19	0.111		-7.9 (-15.6 to 0.5)		-7.4 (-14.5 to 0.3)		-11.6 (-17.9 to -4.7)		-12.1 (-18.4 to -5.5)
	ins	C	17	0.091		-12.3 (-19.7 to -4.1)		-11.3 (-18.2 to -3.8)		-12.8 (-19.3 to -5.8)		-13.3 (-19.7 to -6.4)
COMT	36	324										
	T	A		0.525	0.024	-	0.019	-	0.055		0.088	
	C	G		0.455		-4.9 (-9.7 to 0.1)		-4.7 (-9.1 to -0.2)				
	T	G		0.020		-17.3 (-30.1 to -2.2)		-16.6 (-28.6 to -2.5)				
MTHFR	677	1298										
	C	A		0.369	0.033	-	0.001	-	0.012	-	0.002	-
	C	C		0.335		-1.0 (-6.7 to 5.0)		-0.9 (-6.1 to 4.6)		-0.7 (-5.6 to 4.5)		-0.6 (-5.4 to 4.4)
	T	A		0.296		6.8 (0.9 to 13.1)		9.4 (4.2 to 15.2)		7.3 (2.2 to 12.6)		8.8 (3.8 to 14.0)

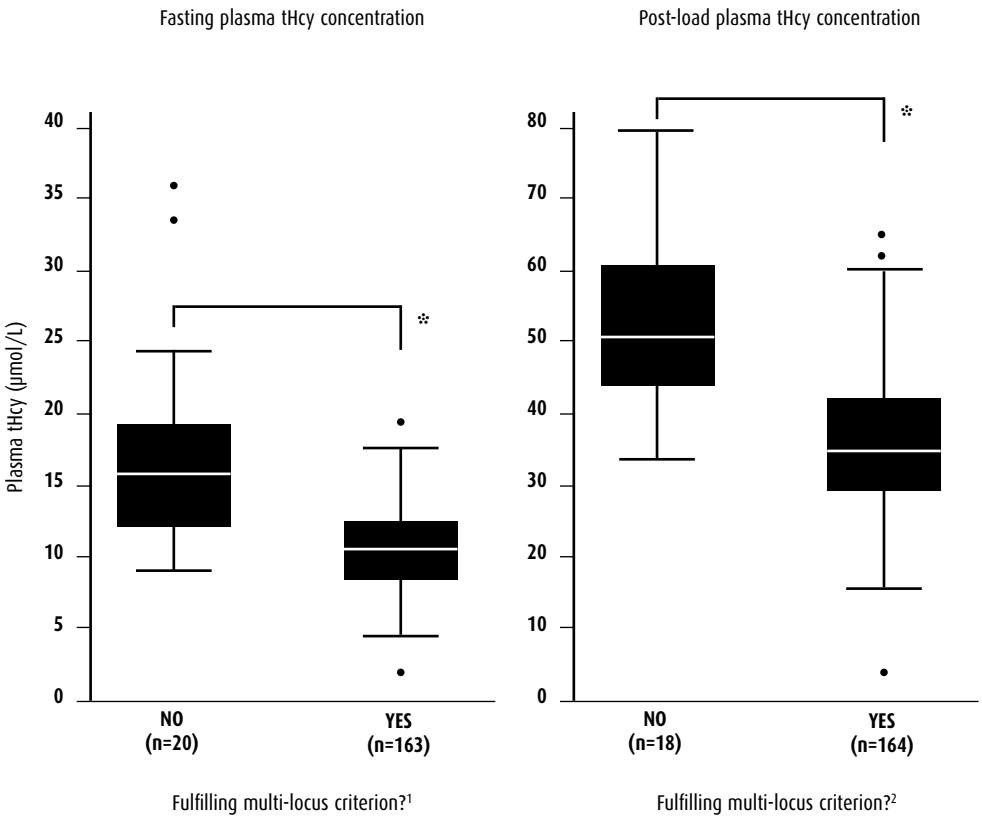

Figure 4.2 *Box plot for unadjusted fasting and post-load plasma tHcy concentration for the group that did and did not suffice the Boolean multi-locus genotype criterion as identified by logic regression. Number of subjects fulfilling the criterion are given in brackets.*

[1] Boolean multi-locus criterion for unadjusted fasting plasma tHcy: [*CBS*31bpVNTR 17-18 or *FOLH1*1561CT/TT or (*BHMT*595GG and *MTHFD*2011AA/AG) or *MTHFR*677CC]

[2] Boolean multi-locus criterion for unadjusted post-load plasma tHcy: [*CBS*31bpVNTR 17-18 or *FOLH1*1561CT/TT or *MTHFD*2011AA/AG or *MTHFR*677CC]

* *P*-value ANOVA < 0.001

Logic regression

Fourteen DNA variants (*BHMT*595G>A, *CBS*699C>T, *CBS*844ins68, *CBS*31bpVNTR, *COMT*324G>A, *DHFR*146_164del, *GCLM*-588C>T, *FOLH1*1561C>T, *UCP2*1032-1077dup, *MTHFD*2011G>A, *MTHFR*677C>T, *MTHFR*1298A>C, *SHMT1*1420C>T, *SLC19A1*80G>A) and 183 and 182 complete observations for fasting and post-load plasma tHcy, respectively, were included in the logic regression analysis. Single locus results for unadjusted fast-

ing plasma tHcy with nominal P<0.05 in this subpopulation included *BHMT*595G>A, *CBS*844ins68, *DHFR*146_164del, *FOLH1*1561C>T, *MTHFR*677C>T, and *UCP2*1032-1077dup. For post-load plasma tHcy *CBS*699C>T, *CBS*844ins68, *CBS* 31bp VNTR, *MTHFD*2011A>G, *MTHFR*677C>T, *SLC19A1*80G>A, *SHMT*1420C>T, and *UCP2*1032-1077dup were found.

The identified logic tree variables for fasting and post-load plasma tHcy are described in Figure 4.2. The *FOLH1*1561C>T, *MTHFD*2011G>A, and *MTHFR*677C>T were consistently included in the Boolean variables. Other included DNA variants were *BHMT*595G>A, and the 31bp VNTR and 844_845ins68 in the *CBS* gene; the latter two variants showed strong LD (r^2>0.8). So not all variables that showed consistent single locus association were found here (e.g. *UCP2*1032-1077dup) and variants that did not show strong marginal association to plasma tHcy were included (e.g. *FOLH1*1561C>T). The identified logic regression models for unadjusted and adjusted fasting plasma tHcy only differed with respect to inclusion of either the 31bp VNTR or 844_845ins68 in the *CBS* gene, two strongly correlated variants. The identified logic regression models for unadjusted and adjusted post-load plasma tHcy were identical (data not shown).

The logic regression models included only one tree and divided the study population in one group with high and one group with low plasma tHcy (Fig. 4.2). Median differences in plasma tHcy for the two groups were 5.2 and 16.6 µmol/L for unadjusted fasting and post-load, respectively. The two groups showed, as expected, statistically significant different plasma tHcy concentrations (ANOVA *P*<0.001). The explained variance in unadjusted fasting, adjusted fasting, unadjusted post-load and adjusted post-load plasma tHcy was 18%, 17%, 16% and 15%, respectively, which was much more than the sum of the marginal contributions (Table 4.4). ANOVA showed that for unad-

Table 4.4 *Percentage explained variance (adjusted R^2) of plasma tHcy concentration by DNA variants based on coding and subsample used in logic regression (n=183 and 182 for fasting and post-load tHcy, respectively).*

DNA variant	Fasting	Fasting adjusted	Post-load	Post-load adjusted
*BHMT*595G>A	1.6	0.7	-	-
*CBS*844_845ins68	-	1.1	-	-
CBS 31bp VNTR	1.1	-	1.6	1.1
*FOLH1*1561C>T	0.0	0.0	0.0	0.0
*MTHFD*2011G>A	0.8	0.4	2.6	2.2
*MTHFR*677C>T	1.3	2.0	2.1	3.2
all DNA variants	4.8	4.5	6.6	6.7
all with pair-wise interactions	16.0	17.0	16.7	16.9
logic regression Boolean variable	17.5	17.4	16.4	15.2

justed fasting plasma tHcy statistically significant interactions between the 31 bp VNTR in *CBS* and *MTHFD*2011G>A, 31 bp VNTR in *CBS* and *BHMT*595G>A, *FOLH1*1561C>T and *MTHFD*2011G>A, *BHMT*595G>A and *MTHFR*677C>T, and *MTHFD*2011G>A and *MTHFR*677C>T were observed in addition to marginal effects of *MTHFR*677C>T and the 31 bp VNTR in *CBS*. For unadjusted post-load plasma tHcy we found statistically significant interactive effects between *MTHFD*2011G>A and *MTHFR*677C>T, *FOLH1*1561C>T and *MTHFD*2011G>A, and *FOLH1*1561C>T and the 31 bp VNTR in *CBS* and no statistically significant marginal effects.

Discussion

In the numerous papers that have described the associations between DNA variants in candidate genes and plasma tHcy in the past, relatively few efforts have been directed towards characterization of multi-locus effects[14]. Factors that may have inhibited these types of analyses are the requirement of genotyping of multiple DNA variants, the need for large sample sizes, and the absence or unfamiliarity with efficient techniques of analysis. However, the application of a haplotype or multi-locus genotype approach offers potential advantages over the single-locus approach in the elucidation of the genetic aetiology of multi-factorial traits. Tools to perform these types of analyses are now widely available[1,2]. Therefore, we re-examined the 36 DNA variants that we have measured over the last 13 years using haplotype and logic regression analysis. We found no haplotype associations independent of the single locus effects. Logic regression analysis learned that a frequent multi-locus genotype was associated with a median increase in fasting plasma tHcy concentrations of ~5 µmol/L.

The results for the separate DNA variants measured in this population were published and discussed by us previously. The small differences between the association results of the current and previous studies can be attributed to differences in number of genotyped individuals and/or approach in statistical analysis. The simultaneous presentation of all single locus results showed that these were dominated by *MTHFR*677C>T and variants in *CBS*, especially for post-load plasma tHcy, but also several other DNA variants contributed to tHcy variation, although less convincing. This is in line with findings from other studies[14]. The *MTHFR*677C>T SNP was the most important marginal contributor to variation in plasma tHcy in our population due to its large effect and common allele frequency of 30% but the explained variance was still limited to ~2%. The explained variance by *MTHFR*677C>T varies among studies; e.g., estimates from just under 2%[48] and over 12%[49] have been reported. These differences may be attributed to the interpopulation variety in *MTHFR*677C>T frequency[50] and other sources of genetic or environmental heterogeneity between these populations.

Our study confirmed the absence of the *MTHFR*677T and *MTHFR*1298C allele on one chromosome which indicates tight physical linkage between the loci or, possibly, fetal non-viability[51,52]. Also, the measured variants in *CBS* and *COMT* showed lack of recombination and were located in one block with limited haplotype diversity. The presence of high LD within *CBS* has been reported for a subset of our population previously[34] and by others[53,54]. Interestingly, we found that the insertion allele was strongly linked with the 17 repeat allele of the *CBS* 31 bp VNTR (D' 0.95, r^2 0.83). However, Vyletal et al.[54], also found haplotypes containing the 844_845ins and the 18 and 21 allele in a Czech population. The effect of the two DNA variants on plasma tHcy was difficult to disentangle in our population due to the high correlation. Previous studies have indicated aberrant splicing and differences in CBS enzyme activity for different number of repeat units for *CBS* 31 bp VNTR[30], and no influence of the *CBS*844_845ins(68bp) on the size of the mRNA product[55] which supports a causal role for the first, and the indirect measurement of this effect by the latter DNA variant. A high correlation between *COMT*324A>G and *COMT*36T>C SNP was found; conditional analysis and known functionality indicated causal importance of 324G>A, as we already reported[20].

The haplotype analysis did not lead to identification of haplotype effects *on top of* single locus effects. This points towards the absence of stronger LD between haplotypes and unmeasured causal variants compared to the single DNA variants, which is very likely related to our strategy of selecting functional variants, and the absence of allelic interactions between the measured DNA variants[2].

A multi-locus genotype based on *BHMT*595G>A, the 31bp VNTR (or 844_845ins68) in the *CBS* gene, *FOLH1*1561C>T, *MTHFD*2011G>A, and *MTHFR*677C>T distinguished two groups with large differences in fasting and post-load plasma tHcy. The deleterious multi-locus genotype was found in 10% of our study subjects. It explained 17% of the variation in plasma tHcy in our population, largely via statistical interactions, and, to a lesser extent, via marginal effects of *MTHFR*677C>T and *CBS* 31 bp VNTR. This underlines the importance of the *combination* of DNA variants in maintaining adequate levels of plasma tHcy. It would be very interesting to study whether this subgroup, that is genetically predisposed to high plasma tHcy, is also at higher risk for developing tHcy-associated disease states.

We previously reported the interaction between *ATIC*346C>G and *SLC19A1*80G>A[19] and 31 bp VNTR in *CBS* and *MTHFR*677C>T[16] based on traditional regression analysis in our study population. The latter two variants were also included in the multi-locus genotype in the current study. Our logic regression analysis also confirmed the modifying effect of the *CBS*844_845ins68 on plasma tHcy increase in *MTHFR*677TT homozygotes as recently reported by Summers et al.[56]. Other studies, that have investigated combined genotype effects using traditional regression methods, have reported statistical interactions between the *MTHFR*677C>T and *SLC19A*80A>G[51], *MTHFR*677C>T and *MTR*2576A>G and *MTRR*66A>G[57], and *MTHFR*677C>T and *MTHFR*1298A>C,

*SLC19A*80A>G, *MTR*2576A>G, and *MTR*R66A>G[58]. These genotype combinations were not part of the multi-locus genotype identified in the current study. Post-hoc analysis in our population using traditional regression analysis showed statistical interaction between *MTHFR*677C>T and *MTRR*66A>G for unadjusted fasting plasma tHcy only (data not shown).

A strong point in our study is the measurement of a large number of (potentially) functional DNA variants in several key genes in homocysteine metabolism. This increased our prior chance of finding biologically meaningful genetic determinants of plasma tHcy. We have most likely not captured *all* genetic variation in the chosen candidate genes which makes it impossible to make final statements about involvement on gene level. Future studies would benefit from inclusion of additional DNA variants to cover all variation within the genes.

A potential source of bias in gene-association studies is presence of population stratification. Our study population was recruited from a small geographic area in The Hague and we only included individuals from known Caucasian ancestry, thereby limiting the chance of population stratification bias. In addition to crude analyses, we performed analyses for age, sex, and creatinine adjusted plasma tHcy levels. The standardized measurement of fasting as well as post-load plasma tHcy minimized the variation due to short-term dietary intake and allowed for comparison of the genetic associations for two homocysteine phenotypes.

Our study population at large consisted of 500 individuals which gave us 80% power to detect a statistically significant association for DNA variants that explained 1.5% of the trait variance at an alpha level of 0.05. Unfortunately, fewer subjects were analysed due to missing genotype information which limited our power (e.g. 300 individuals gave 80% power to detect a DNA variant with R^2 2.5% at alpha 0.05). Missing values also hampered logic regression analysis. To increase the number of included individuals, we omitted DNA variants that showed no marginal effect on plasma tHcy, thereby introducing a bias against these DNA variants. External validation studies will need to verify whether the large effect on plasma tHcy concentration of the identified multi-locus genotype in our population can be generalized.

Our systematic multi-locus association analysis for a large number of candidate gene variants allowed us to characterize LD patterns, demonstrate the absence of allelic interactions between the measured gene variants and the absence of strong LD between the estimated haplotypes and unmeasured causal variants. Moreover, our analysis approach led to the identification of a multi-locus genotype that divided our population in two subgroups that showed over 5 µmol/L difference in median fasting plasma tHcy. The presence of interactions and small marginal effects of DNA variants, selected for their high prior probability of involvement in plasma tHcy level, confirmed the complex character of plasma tHcy and underlines the importance of multi-locus analysis in the elucidation of its genetic background.

Acknowledgments

This work was supported by the Netherlands Heart Foundation, Grant 2002B68. Sandra Heil was supported by Grant C042083 of the Dutch Kidney Foundation. Martin den Heijer received a VENI grant from the Netherlands Organization for Scientific Research (NWO).

References

1. Hoh J, Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet. 2003;4:701-709.
2. Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol. 2004;27:348-364.
3. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. BMJ. 2002;325:1202.
4. The Homocysteine Studies Collaboration. Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA. 2002;288:2015-2022.
5. Casas JP, Bautista LE, Smeeth L, Sharma P, Hingorani AD. Homocysteine and stroke: evidence on a causal link from mendelian randomisation. Lancet. 2005;365:224-232.
6. den Heijer M, Lewington S, Clarke R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost. 2005;3:292-299.
7. Nelen WL. Hyperhomocysteinaemia and human reproduction. Clin Chem Lab Med. 2001;39:758-763.
8. Morris MS. Homocysteine and Alzheimer's disease. Lancet Neurol. 2003;2:425-428.
9. Bazzano LA, Reynolds K, Holder KN, He J. Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials. JAMA. 2006;296:2720-2726.
10. B-Vitamin Treatment Trialists' Collaboration. Homocysteine-lowering trials for prevention of cardiovascular events: a review of the design and power of the large randomized trials. Am Heart J. 2006;151:282-287.
11. Wang X, Qin X, Demirtas H, Li J, Mao G, Huo Y, Sun N, Liu L, Xu X. Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. Lancet. 2007;369:1876-1882.
12. De Bree A, Verschuren WM, Kromhout D, Kluijtmans LA, Blom HJ. Homocysteine determinants and the evidence to what extent homocysteine determines the risk of coronary heart disease. Pharmacol Rev. 2002;54:599-618.
13. den Heijer M, Graafsma S, Lee SY, van Landeghem B, Kluijtmans L, Verhoef P, Beaty TH, Blom H. Homocysteine levels--before and after methionine loading--in 51 Dutch families. Eur J Hum Genet. 2005;13:753-762.

14. Gellekink H, den Heijer M, Heil SG, Blom HJ. Genetic determinants of plasma total homocysteine. Semin Vasc Med. 2005;5:98-109.
15. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJH, den Heijer M, Kluijtmans LAJ, van den Heuvel LP, Rozen R. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. Nat Genet. 1995;10:111-113.
16. Afman LA, Lievers KJ, Kluijtmans LA, Trijbels FJ, Blom HJ. Gene-gene interaction between the cystathionine beta-synthase 31 base pair variable number of tandem repeats and the methylenetetrahydrofolate reductase 677C > T polymorphism on homocysteine levels and risk for neural tube defects. Mol Genet Metab. 2003;78:211-215.
17. Gellekink H, den Heijer M, Kluijtmans LA, Blom HJ. Effect of genetic variation in the human S-adenosylhomocysteine hydrolase gene on total homocysteine concentrations and risk of recurrent venous thrombosis. Eur J Hum Genet. 2004;12:942-948.
18. Gellekink H, Blom HJ, van der Linden I, den Heijer M. Molecular genetic analysis of the human dihydrofolate reductase gene: relation with plasma total homocysteine, serum and red blood cell folate levels. Eur J Hum Genet. 2007;15:103-109.
19. Gellekink H, Blom HJ, den Heijer M. Associations of common polymorphisms in the thymidylate synthase, reduced folate carrier and 5-aminoimidazole-4-carboxamide ribonucleotide transformylase/inosine monophosphate cyclohydrolase genes with folate and homocysteine levels and venous thrombosis risk. Clin Chem Lab Med. 2007;45:471-476.
20. Gellekink H, Muntjewerff JW, Vermeulen SH, Hermus AR, Blom HJ, den Heijer M. Catechol-O-methyltransferase genotype is associated with plasma total homocysteine levels and may increase venous thrombosis risk. Thromb Haemost. 2007;98:1226-1231.
21. Heil SG, Lievers KJ, Boers GH, Verhoef P, den Heijer M, Trijbels FJ, Blom HJ. Betaine-homocysteine methyltransferase (BHMT): genomic sequencing and relevance to hyperhomocysteinemia and vascular disease in humans. Mol Genet Metab. 2000;71:511-519.
22. Heil SG, van der Put NM, Waas ET, den Heijer M, Trijbels FJ, Blom HJ. Is mutated serine hydroxymethyltransferase (SHMT) involved in the etiology of neural tube defects? Mol Genet Metab. 2001;73:164-172.
23. Heil SG, den Heijer M, Van Der Rijt-Pisa BJ, Kluijtmans LA, Blom HJ. The 894 G > T variant of endothelial nitric oxide synthase (eNOS) increases the risk of recurrent venous thrombosis through interaction with elevated homocysteine levels. J Thromb Haemost. 2004;2:750-753.
24. Heil SG, Kluijtmans LA, De Vriese AS, den Heijer M, Blom HJ. Evidence for a critical role of the modifying subunit of glutamate-cysteine ligase in homocysteine pathophysiology. In: Heil SG: Unraveling the mystery of homocysteine. Radboud University Nijmegen; 2006. pp 99-112.
25. Heil SG, Vermeulen SH, Van der Rijt-Pisa BJ, den Heijer M, Blom HJ. Role for mitochondrial uncoupling protein-2 (UCP2) in hyperhomocysteinemia and venous thrombosis risk? Clin Chem Lab Med. 2008;46:655-659.

26. Hol FA, van der Put NM, Geurds MP, Heil SG, Trijbels FJ, Hamel BC, Mariman ECM, Blom HJ. Molecular genetic analysis of the gene encoding the trifunctional enzyme MTHFD (methylenetetrahydrofolate-dehydrogenase, methenyltetrahydrofolate-cyclohydrolase, formyltetrahydrofolate synthetase) in patients with neural tube defects. Clin Genet. 1998;53:119-125.
27. Klerk M, Lievers KJ, Kluijtmans LA, Blom HJ, den Heijer M, Schouten EG, Kok FJ, Verhoef P. The 2756A>G variant in the gene encoding methionine synthase: its relation with plasma homocysteine levels and risk of coronary heart disease in a Dutch case-control study. Thromb Res. 2003;110:87-91.
28. Kluijtmans LA, van den Heuvel LP, Boers GH, Frosst P, Stevens EM, van Oost BA, den Heijer M, Trijbels FJ, Rozen R, Blom HJ. Molecular genetic analysis in mild hyperhomocysteinemia: a common mutation in the methylenetetrahydrofolate reductase gene is a genetic risk factor for cardiovascular disease. Am J Hum Genet. 1996;58:35-41.
29. Kluijtmans LA, Boers GH, Trijbels FJ, van Lith-Zanders HM, van den Heuvel LP, Blom HJ. A common 844INS68 insertion variant in the cystathionine beta-synthase gene. Biochem Mol Med. 1997;62:23-25.
30. Lievers KJ, Kluijtmans LA, Heil SG, Boers GH, Verhoef P, van Oppenraay-Emmerzaal D, den Heijer M. Trijbels FJ, Blom HJ. A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. Eur J Hum Genet. 2001;9:583-589.
31. Lievers KJ, Boers GH, Verhoef P, den Heijer M, Kluijtmans LA, van der Put NM, Trijbels FJ, Blom HJ. A second common variant in the methylenetetrahydrofolate reductase (MTHFR) gene and its relationship to MTHFR enzyme activity, homocysteine, and cardiovascular disease risk. J Mol Med. 2001;79:522-528.
32. Lievers KJ, Afman LA, Kluijtmans LA, Boers GH, Verhoef P, den Heijer M, Trijbels FJ, Blom HJ. Polymorphisms in the transcobalamin gene: association with plasma homocysteine in healthy individuals and vascular disease patients. Clin Chem. 2002;48:1383-1389.
33. Lievers KJ, Kluijtmans LA, Boers GH, Verhoef P, den Heijer M, Trijbels FJ, Blom HJ. Influence of a glutamate carboxypeptidase II (GCPII) polymorphism (1561C-->T) on plasma homocysteine, folate and vitamin B(12) levels and its relationship to cardiovascular disease risk. Atherosclerosis. 2002;164:269-273.
34. Lievers KJ, Kluijtmans LA, Heil SG, Boers GH, Verhoef P, Den Heijer M, Trijbels FJ, Blom HJ. Cystathionine beta-synthase polymorphisms and hyperhomocysteinaemia: an association study. Eur J Hum Genet. 2003;11:23-29.
35. van der Linden I, den Heijer M, Afman LA, Gellekink H, Vermeulen SH, Kluijtmans LA, Blom HJ. The methionine synthase reductase 66A>G polymorphism is a maternal risk factor for spina bifida. J Mol Med. 2006;84:1047-1054.
36. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet. 1995;346:1070-1071.

37. den Heijer M, Blom HJ, Gerrits WB. Is hyperhomocysteinaemia a risk factor for recurrent venous thrombosis? Lancet. 1995;345:882-885.
38. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16:1215.
39. den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet. 2001;109:121-124.
40. te Poele-Pothoff MT, van den Berg M, Franken DG, Boers GH, Jakobs C, de Kroon IF, Eskes TK, Trijbels JM, Blom HJ. Three different methods for the determination of total homocysteine in plasma. Ann Clin Biochem. 1995;32:218-220.
41. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 1992;48:361-372.
42. Cleves MA. Hardy-Weinberg Equilibrium Tests and Allele Frequency Estimation. STATA Technical Bulletin. 1999;48:34-37.
43. Gabriel SB, Schaffner SF, Nguyen H. The structure of haplotype blocks in the human genome. Science. 2002;296:2225-2229.
44. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263-265.
45. Purcell S, Daly MJ, Sham PC. WHAP: haplotype-based association analysis. Bioinformatics. 2007;23:255-256.
46. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet Epidemiol. 2001;21 Suppl 1:S626-S631.
47. Ruczinski I, Kooperberg C, LeBlanc L. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression, with applications. J Multiv Anal. 2004;90:178-195.
48. Dekou V, Gudnason V, Hawe E, Miller GJ, Stansbie D, Humphries SE. Gene-environment and gene-gene interaction in the determination of plasma homocysteine levels in healthy middle-aged men. Thromb Haemost. 2001;85:67-74.
49. Gudnason V, Stansbie D, Scott J, Bowron A, Nicaud V, Humphries S. C677T (thermolabile alanine/valine) polymorphism in methylenetetrahydrofolate reductase (MTHFR): its frequency and impact on plasma homocysteine concentration in different European populations. EARS group. Atherosclerosis. 1998;136:347-354.
50. Botto LD, Yang Q. 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. Am J Epidemiol. 2000;151:862-877.
51. Devlin AM, Clarke R, Birks J, Evans JG, Halsted CH. Interactions among polymorphisms in folate-metabolizing genes and serum total homocysteine concentrations in a healthy elderly population. Am J Clin Nutr. 2006;83:708-713.
52. Ogino S, Wilson RB. Genotype and haplotype distributions of MTHFR677C>T and 1298A>C single nucleotide polymorphisms: a meta-analysis. J Hum Genet. 2003;48:1-7.
53. De Stefano V, Dekou V, Nicaud V, Chasse JF, London J, Stansbie D, Humphries SE, Gudnason V. Linkage disequilibrium at the cystathionine beta synthase (CBS) locus and the association between genetic variation at the CBS locus and plasma levels of homocysteine. The Ears II Group. European Atherosclerosis Research Study. Ann Hum Genet. 1998;62:481-490.

54. Vyletal P, Sokolová J, Cooper DN, Kraus JP, Krawczak M, Pepe G, Rickards O, Koch HG, Linnebank M, Kluijtmans LA, Blom HJ, Boers GH, Gaustadnes M, Skovby F, Wilcken B, Wilcken DE, Andria G, Sebastio G, Naughten ER, Yap S, Ohura T, Pronicka E, Laszlo A, Kozich V. Diversity of cystathionine beta-synthase haplotypes bearing the most common homocystinuria mutation c.833T>C: a possible role for gene conversion. Hum Mutat. 2007;28:255-264.
55. Tsai MY, Bignell M, Schwichtenberg K, Hanson NQ. High prevalence of a mutation in the cystathionine beta-synthase gene. Am J Hum Genet. 1996;59:1262-1267.
56. Summers CM, Hammons AL, Mitchell LE, Woodside JV, Yarnell JW, Young IS, Evans A, Whitehead AS. Influence of the cystathionine ß-synthase 844ins68 and methylenetetrahydrofolate reductase 677C>T polymorphisms on folate and homocysteine concentrations. Eur J Hum Genet. 2008;16:1010-1013.
57. Kluijtmans LA, Young IS, Boreham CA, Murray L, McMaster D, McNulty H, Strain JJ, McPartlin J, Scott JM, Whitehead AS. Genetic and nutritional factors contributing to hyperhomocysteinemia in young adults. Blood. 2003;101:2483-2488.
58. Lucock M, Yates Z. Synergy between 677 TT MTHFR genotype and related folate SNPs regulates homocysteine level. Nutrition Research. 2006;26:180-185.

Sita HHM Vermeulen*
Barbara Franke*
Regine PM Steegers-Theunissen
Marieke J Coenen
Mascha MVAP Schijvenaars
Hans Scheffer
Martin den Heijer
Henk J Blom

*These authors contributed equally to this work.

Published in: Birth Defects Research A: Clinical and Molecular Teratology 2009;85:216-226

CHAPTER 5

An association study of 45 folate-related genes in spina bifida: Involvement of cubilin (*CUBN*) and tRNA aspartic acid methyltransferase 1 (*TRDMT1*)

Abstract

Background:
Spina bifida is a class of neural tube defects, which are congenital malformations of the central nervous system with a prevalence of 0.5 to 12 per 1000 births globally. In this article we attempt to identify genes related to folate and its metabolic pathways that are involved in the etiology of spina bifida.

Methods:
We selected 50 folate metabolism-related genes and genotyped polymorphisms in those genes. Eighty-seven polymorphisms in 45 genes passed quality controls. Associations with spina bifida were investigated in 180 patients and 190 controls. For those polymorphisms that were nominally associated with spina bifida risk, the relation with serum and red blood cell folate, vitamin B_{12}, and homocysteine was evaluated in controls.

Results:
A polymorphism in *CUBN* was significantly associated with decreased spina bifida risk, after correction for multiple testing, and was related to increased vitamin B_{12} ($p = 0.039$) and red blood cell folate ($p = 0.001$). The *CUBN* gene encodes the intrinsic factor-cobalamin receptor (or cubilin), a peripheral membrane protein that acts as a receptor for intrinsic factor-vitamin B_{12} complexes. Vitamin B_{12} is an important cofactor in the folate metabolism, and low B_{12} status in mothers has been linked to neural tube defects in children. Other interesting findings include nominally significant associations with polymorphisms in *TRDMT1*, *ALDH1L1*, *SARDH*, and *SLCA19A1* (*RFC1*).

Conclusion:
Our study indicates interesting new candidate genes and functional pathways for further study and confirms earlier findings. None of the genes *CUBN*, *TRDMT1*, *ALDH1L1*, or *SARDH* have been investigated previously for association with spina bifida.

Introduction

Neural tube defects (NTDs) are congenital malformations of the central nervous system with a prevalence of 0.5 to 12 per 1000 births globally. Spina bifida and anencephaly are the two most common types of NTD. Spina bifida is due to a closure defect of the caudal part of the neural tube and is compatible with life. In contrast, cranial closure defects of the neural tube result in lethal anencephaly. Neural tube defects belong to the group of multifactorial disorders caused by genetic predisposition in the presence of unfavorable environmental factors.

Low maternal intake of folate was identified as an important environmental risk factor for NTDs, leading to preventive measures such as periconceptional folic acid supplementation and stimulation of food fortification with folic acid. In this way the occurrence and recurrence of NTDs can be reduced by 50 to 85%[1,2,3]. Metabolites of folate play an important role as cofactors of many different enzymes involved in processes such as purine and pyrimidine synthesis and DNA- and protein-methylation. Given the profound effect of food fortification on neural tube closure in the embryo, it is plausible that the efficiency of the endogenous folate metabolism is an important determinant of NTD risk. The fact that vitamin B_{12} (cobalamin) and homocysteine levels, two biochemical parameters involved in folate metabolism, have also been associated with NTD risk underscores the importance of this metabolism in NTD etiology[4,5,6]. Genetic variation causing suboptimal functioning of the endogenous folate metabolism likely forms the explanation for the fact that even today, in women with normal folate levels, supplementary folic acid can still reduce the risk of having a child with an NTD.

Indeed, the 677C>T polymorphism in the methylenetetrahydrofolate reductase (MTHFR) gene is associated with an increased risk for NTD[7]. A few additional NTD risk factors have been identified in other folate-related genes. However, a large part of the genetic contribution to NTDs still remains unexplained[7,8]. In this article we aim to identify genes related to folate and its metabolic pathways that are involved in the etiology of spina bifida. In addition, for selected polymorphisms, the relation with serum and red blood cell folate, vitamin B_{12}, and homocysteine levels is evaluated in the control sample.

Materials and methods

Study Sample

With the approval of central and local ethics committees, patients with nonsyndromic spina bifida were recruited from the Dutch population between 1989 and 2001. All 180 patients with spina bifida aperta included in this study have been described before in publications of the participating departments[6,9-11]. No information on folic acid

supplementation during pregnancy by case mothers was present for the samples collected before 1992, but we can assume that mothers did not use supplementation at that time. Supplementation started only in 1992 in the Netherlands. For the remaining cases, data on vitamin supplementation by mothers were collected via questionnaires; 36 mothers had used supplements containing folic acid during pregnancy, 43 mothers had not used any vitamin supplement, and data for five mothers were missing. Because only a few mothers had supplementation and dose-information was not available, we performed our analyses without evaluation of the potential effect modification by folic acid supplementation.

One hundred and ninety subjects of Dutch ethnicity were selected randomly from a group consisting of 500 unrelated healthy subjects that was ascertained via a general practice in The Hague[12]. The characteristics of this population and the measurements of serum and red blood cell (RBC) folate, serum vitamin B_{12}, and fasting plasma total homocysteine (tHcy) have been described previously[12].

Selection of Candidate Genes and Polymorphisms

Candidate genes and polymorphisms were selected from literature and databases[13,14] during 2002. We preferentially chose those polymorphisms that were either functional (causing changes in gene function or regulation) or had already shown association with any disease or condition. From databases, we preferentially selected single nucleotide polymorphisms (SNPs) that had been confirmed, had a known frequency in Caucasians, and were potentially functional or within or near exons or the promoter region. In genes with known functional polymorphisms, we did not select additional variants for analysis.

Genotyping

Genotyping of most SNPs was performed at ASPER Biotech (Tartu, Estonia) using arrayed primer extension (APEX) technology [15]. Genotyping was performed in our own departments if APEX assay-design for an SNP was not possible and for all non-SNP polymorphisms, using restriction fragment length polymorphism analysis[6] or by direct determination of allele length. Assay conditions for individual polymorphisms genotyped in Nijmegen, The Netherlands, are available from the corresponding author. Genotyping of 154 polymorphisms (146 SNPs and 8 repeat polymorphisms) was attempted. As controls for genotyping, 5% blind duplicates and blanks were included in all sample plates. In addition, in selected samples eight SNPs genotyped by ASPER were genotyped by a second technique to confirm assay integrity. For quality control reasons, we included only those polymorphisms in the analysis that (1) had a minor allele frequency higher than 0.02, (2) were in Hardy-Weinberg equilibrium (p value >0.01), and (3) had less than 25% missing genotypes. Of the 154 polymorphisms in 50 genes selected for analysis, 87 polymorphisms in 45 genes passed the quality controls.

Statistical Analysis

Test for deviation from Hardy-Weinberg equilibrium was performed with chi-square or Fisher's exact tests in controls using the GENHW add-on package[16] or a Monte Carlo exact test[17]. Genotypic chi-square tests for association of the diallelic sequence variants with case-control status were performed assuming a dominant, recessive, and a genotypic model. Odds ratios (ORs) and 95% confidence intervals (CI) were constructed using logistic regression. Multiallelic markers were recoded into diallelic variants (Table 5.1) and analyzed as described previously. Nominal (i.e., uncorrected for multiple comparisons) two-tailed p values are presented. Correction for multiple comparisons according to the false discovery rate controlling procedure[19] was applied with a q-value setting of 0.05 using the multproc add-on package developed by Newson[20]. Given our sample size, this study has a 78% power to find with nominal significance (p value <0.05) a genetic variant with an OR of 1.5 and a minor allele requency of 40%. The relation between the nominally associated polymorphisms and age and sex adjusted serum and RBC folate, vitamin B_{12}, and tHcy levels was evaluated using linear regression. The distributions of the biochemical parameters were skewed to the right and therefore logarithmically transformed before analysis. All analyses were performed in Stata version 9.0 (StataCorp LP, College Station, TX).

Haplotype analyses were conducted only for those genes that showed suggestive nominal association in the single locus chi-square analysis (nominal p value <0.10). Pairwise linkage disequilibrium (LD; D and r^2) for diallelic sequence variants located in one gene was computed using Haploview (Broad Institute, Cambridge, MA). Haplotype association analysis was performed (for haplotypes with a frequency 2%) using Whap[21]. Two strategies were followed: haplotype analysis for identified blocks and sliding window approach for 2, 3, and 4 consecutive sequence variants in a gene. We used 1000 permutations to generate empirical p values in Whap.

Explorative analysis of gene-gene interactions in relation to risk of NTD was performed using the logic regression package in R (version 2.3.1)[22] for those sequence variants that showed single locus associations with a nominal p value <0.10. Ten-fold cross-validation was performed to identify the best model.

Results

Table 5.1 describes the genes and polymorphisms that were selected for analysis. Genotype frequencies, ORs, and 95% CIs for the six polymorphisms that displayed a nominal chi-square p value <0.05 for association with spina bifida risk are displayed in Table 5.2. These nominally significant findings included two uncorrelated (r^2 = 0.003) polymorphisms in *CUBN*, and one in *TRDMT1* (=*DNMT2*), *ALDH1L1*, *SARDH*, and *SLC19A1* (=*RFC1*). After correction for multiple testing using the false discovery rate approach, only rs1907362 in *CUBN* was statistically significantly associated with disease status (critical p value 0.0004065).

Table 5.1 *Description of genes and polymorphisms selected for analysis.*

Gene symbol[a]	Gene name[a]	Contig accession	Folate-related process[b]	Chromosomal location	Size gene in kb[c]	Gene variant[d]	Protein change	Genotyping technique	Reason exclusion [e]
AHCY	S-adenosylhomocysteine hydrolase	NT_028392.5	3	20cen-q13.1	23.1	rs1205366		APEX	
AHCY						rs819158		APEX	
AHCY						rs6088457	Lys318Lys	APEX	1,2
AHCY						rs6058020		APEX	1,2
ALDH1L1	aldehyde dehydrogenase 1 family, member L1	NT_005612.14	2	3q21.3	77.1	rs4646696		APEX	2,3
ALDH1L1						rs1868138		APEX	
ALDH1L1						rs3796191	Leu254Pro	APEX	
ALDH1L1						rs2305230	Leu395Leu	APEX	
ALDH1L1						rs873696		APEX	
ALDH1L1						**rs2290053**[f]		APEX	
ALDH1L1						**rs1127717**	Asp793Gly	APEX	
AMD1	adenosylmethionine decarboxylase 1	NT_025741.13	6	6q21-q22	20.9	rs1049705	Leu29Leu	APEX	1,2
AMD1						rs1049699		APEX	2
AMD1						rs1007274		RFLP analysis	3
AMD1						rs989651		APEX	2
AMT	aminomethyltransferase	NT_022517.17	5	3p21.2-p21.1	5.8	rs10640		APEX	
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	NT_005403.15	7	2q35	37.6	rs4673991		APEX	
ATIC						rs1997059		APEX	
ATIC						rs2372536	Thr116Ser	APEX	
ATIC						rs4672766		APEX	2,3
BHMT	betaine homocysteine methyltransferase	NT_006713.14	3	5q13	20.4	rs1050825	Gly372Cys	APEX	1,2
BHMT						rs3733890	Arg239Gln	RFLP analysis	
BHMT						rs672346	Phe219Leu	APEX	
BHMT						**rs651852**		APEX	
BHMT2	betaine homocysteine methyltransferase 2	NT_006713.14	3	5q13.1-q13.2	19.7	rs526264		APEX	

BHMT2						rs682985	Asp54Asp	APEX	
BHMT2						rs670220		APEX	2
CBS	cystathionine-beta-synthase	NT_030188.4	4	21q22.3	23.1	844_845ins(68bp)[g]		Simple sequence analysis	
CBS						g.14037(31bp)16-21[h,i]		RFLP analysis	3
COQ3	coenzyme Q3 homolog, methyltransferase (S. cerevisiae)	NT_025741.13	5	6q16.2-q16.3	24.7	rs4612169		APEX	1
COQ3						rs7755791		APEX	
COQ3						rs4144164	Tyr329His	APEX	1,2
CTH	cystathionase (cystathionine gamma-lyase)	NT_032977.7	4	1p31.1	28.3	rs1021737	Ser403Ile	APEX	
CTH						rs3767205		APEX	
CTH						rs663649		APEX	2
CTH						rs501939		APEX	
CTH						**rs490574**		APEX	
CUBN	cubilin (intrinsic factor-cobalamin receptor)	NT_077569.2	1	10p12.31	305.3	rs932640		APEX	
CUBN						**rs4748353**		APEX	
CUBN						rs7905349	His730Tyr	APEX	1,2
CUBN						rs1801227	Tyr1032His	APEX	1,2
CUBN						**rs1907362**		APEX	
CUBN						rs925522		APEX	2,3
CUBN						rs1801231	Pro1559Ser	APEX	
CUBN						rs1276708	Arg1775W	APEX	1,2
CUBN						rs2271462	Gly1840Ser	APEX	
CUBN						rs2271461		APEX	
CUBN						rs3740168	Pro2575Arg	APEX	1,2
CUBN						rs2221886		APEX	2
CUBN						rs1801239	Ile2984Val	APEX	
CUBN						rs703075		APEX	

Gene symbol[a]	Gene name[a]	Contig accession	Folate-related process[b]	Chromosomal location	Size gene in kb[c]	Gene variant[d]	Protein change	Genotyping technique	Reason exclusion [e]
CUBN						rs703064	Thr3281Thr	APEX	2
CUBN						rs1801232	Asn3552Lys	APEX	1,2
DHFR	dihydrofolate reductase	NT_006713.14	2	5q11.2-q13.2	28.8	IVS1+261delACCTGGG CGGGACGCGCCA		APEX	
DHFR						rs34764978		APEX	2
DNMT1	DNA (cytosine-5-)-methyltransferase 1	NT_011295.10	5	19p13.3-p13.2	61.8	rs3745269		APEX	2,3
DNMT1						rs2290684		APEX	
DNMT1						rs8101626		APEX	
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	NT_022184.14	5	2p23	109.6	rs1465764		APEX	
DNMT3A						rs4563176		APEX	2
DNMT3A						rs2276598	Leu422Leu	APEX	
DNMT3A						rs2289195		APEX	
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta	NT_028392.5	5	20q11.2	47.0	IVS17-5C>T		APEX	1,2
DNMT3B						rs7354578	Ala204Ser	APEX	1,2
FOLH1	folate hydrolase (prostate-specific membrane antigen) 1	NT_009237.17	1	11p11.2	62.0	1320C>T	His475Tyr	RFLP analysis	
FOLR1	folate receptor 1 (adult)	NT_033927.7	1	11q13.3-q13.5	6.7	rs1801932	Trp160Cys	APEX	1,2
FOLR1						rs7928649	Trp28Arg	APEX	1,2
FOLR1						rs1893007		APEX	
FOLR2	folate receptor 2 (fetal)	NT_033927.7	1	11q13.3-q13.5	5.1	-631T>C		APEX	1,2
FOLR2						-762G>A		APEX	3
FOLR2						IVS1+112(TGTAT)3-8[j]		SSLP analysis	
FOLR3	folate receptor 3 (gamma)	NT_033927.7	1	11q13	4.1	rs1802608	Leu193Phe	APEX	
FOLR3						rs533207		APEX	
FPGS	folylpolyglutamate synthase	NT_008470.17	2	9cen-q34	11.4	rs10106		APEX	2
FPGS						rs7856096		APEX	1,2

FTCD	formiminotransferase cyclodeaminase	NT_011515.11	2	21q22.3	19.3	rs7277617		APEX	
FTCD						rs4819205		APEX	
FTCD						**rs12774**		APEX	
GAMT	guanidinoacetate N-methyltransferase	NT_011255.14	5	19p13.3	4.5	rs2074899		APEX	
GART	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	NT_011512.10	7	21q22.1	38.1	rs1804387	Leu21Phe	APEX	1,2
GART						rs2834235		RFLP analysis	3
GART						rs2834234		APEX	2
GART						rs8971	Asp752Gly	APEX	
GGH	gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	NT_008183.18	2	8q12.1	23.7	rs719235		APEX	
GGH						rs2305558		APEX	2
GGH						rs1031552		APEX	
GNMT	glycine N-methyltransferase	NT_007592.14	5	6p12	3.1	rs2296805		APEX	2
ICMT	isoprenylcysteine carboxyl methyltransferase	NT_021937.17	5	1p36.21	14.8	rs846108		APEX	1,2
ICMT						**rs1802353**		APEX	
MAT1A	methionine adenosyltransferase I, alpha	NT_030059.12	3	10q22	17.9	rs2342812		APEX	
MAT1A						rs756208		APEX	
MAT1A						rs2993763	Tyr377Tyr	APEX	2
MAT2A	methionine adenosyltransferase II, alpha	NT_022184.14	3	2p11.2	6.1	rs1078004	Arg264Arg	APEX	
MAT2A						rs2028106		APEX	2,3
MGMT	O-6-methylguanine-DNA methyltransferase	NT_008818.15	5	10q26.3	299.9	rs2308321	Ile143Val	APEX	
MGMT						rs12917	Leu84Phe	APEX	
MGMT						rs2020893	Glu30Lys	APEX	
MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1	NT_026437.11	2	14q24	71.6	rs1803951	Gly794Cys	APEX	1,2
MTHFD1						rs2236225	Arg653Gln	APEX	

Gene symbol[a]	Gene name[a]	Contig accession	Folate-related process[b]	Chromosomal location	Size gene in kb[c]	Gene variant[d]	Protein change	Genotyping technique	Reason exclusion [e]
MTHFD1						**rs1950902**	Lys134Arg	APEX	
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2	NT_022184.14	2	2p12	16.7	rs12196		APEX	
MTHFD2						rs6758492	Ala62Thr	APEX	1,2
MTHFD2						rs1667627		APEX	
MTHFR	5,10-methylenetetrahydrofolate reductase	NT_021937.17	2	1p36.3	20.3	rs1801133	Ala222Val	APEX	
MTHFR						rs1801131	Ala429Glu	APEX	
MTHFR						rs2274976	Arg594Gln	APEX	
MTHFS	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)	NT_010194.16	2	15q23	52.0	rs8023452		APEX	1,3
MTHFS						rs2586183		APEX	
MTR	5-methyltetrahydrofolate-homocysteine methyltransferase	NT_004836.16	3	1q42.3-q43	105.2	rs1805087	Asp919Gly	APEX	
MTRR	5-methyltetrahydrofolate-homocysteine methyltransferase reductase	NT_006576.15	3	5q15.3-p15.2	32.0	rs1801394	Ile22Met	APEX	
NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	NT_030737.9	8	8p22	10.0	rs1799931	Gly286Glu	APEX	
NAT2						rs1799930	Arg197Gln	APEX	2
NAT2						rs1801280	Ile114Thr	APEX	2
NNMT	nicotinamide N-methyltransferase	NT_033899.7	5	11q23.1	16.7	rs1050207	Thr245Pro	APEX	1,2
NNMT						rs1941404		APEX	
NOS1	nitric oxide synthase 1 (neuronal)	NT_009775.15	8	2q24.2-q24.31	148.6	IVS14+267(AAT)8-16[k]		SSLP analysis	
NOS1						4775(CA)14-22[l]		SSLP analysis	
NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	NT_010799.14	8	17q11.2-q12	43.8	IVS1-2660(CCTTT)8-18[m]		GeneScan analysis	
NOS2A						rs12720460		GeneScan analysis	3

NOS2A						rs2297518	Ser608Leu	RFLP analysis	
NOS3	nitric oxide synthase 3 (endothelial cell)	NT_007914.14	8	7q36.1	23.5	rs1799983	Asp298Glu	RFLP analysis	
NOS3						IVS4+245(GAAGTCTAGACC TGCTGCAGGGGTGAG)4-6^{n}		Simple sequence analysis	
NOS3						rs2070744		APEX	3
PCMT1	protein-L-isoaspartate (D-aspartate) O-methyltransferase	NT_025741.13	5	6q24-25.1	61.6	rs4816	Val120Ile	APEX	3
PCMT1						rs4038682		APEX	3
PRMT1	protein arginine methyltransferase 1	NT_011109.15	5	19q13.33	11.2	rs3745468		APEX	2
PRMT1						rs1128424	Glu108Val	APEX	1,2
PRMT1						rs1804486	Lys78Met	APEX	1,2
PRMT1						**rs975484**		APEX	
PRMT2	protein arginine methyltransferase 2	NT_011515.11	5	21q22.3	29.3	rs2070436		APEX	
PRMT2						rs2839370		APEX	2,3
RNMT	RNA (guanine-7-) methyltransferase	NT_010859.14	5	18p11.23-p11.22	37.9	rs4797810		APEX	
RNMT						rs4797808		APEX	1,2
RNMT						rs1801762	Lys93Glu	APEX	1,2
RNMT						rs2226754		APEX	1,2
SARDH	sarcosine dehydrogenase	NT_035014.4	2	9q34.2	76.4	**rs573904**	Gln50Gln	APEX	
SARDH						rs2427979		APEX	1,2
SARDH						**rs2073815**	His489His	APEX	
SARDH						rs2073817	Arg614His	APEX	
SARDH						rs886016	Met648Val	APEX	2
SARDH						rs7859013		APEX	1,2
SARDH						rs2519123	Val730Val	APEX	2
SARDH						rs7854480		APEX	
SHMT1	serine hydroxymethyltransferase 1 (soluble)	NT_010718.15	2	17p11.2	35.7	rs1979277	Leu474Phe	APEX	
SLC19A1	solute carrier family 19 (folate transporter), member 1	NT_011515.11	1	21q22.3	27.7	**rs1051266**	His27Arg	APEX	

Gene symbol[a]	Gene name[a]	Contig accession	Folate-related process[b]	Chromosomal location	Size gene in kb[c]	Gene variant[d]	Protein change	Genotyping technique	Reason exclusion [e]
SLC19A1						rs12659	Pro114Pro	APEX	2
SLC19A1						rs1051296		APEX	2,3
TCN2	transcobalamin II	NT_011520.10	1	22q11.2-qter	19.9	rs1801198	Arg259Pro	APEX	
TRDMT1	tRNA aspartic acid methyltransferase 1	NT_077569.2	5	10p13	58.7	rs1885396		APEX	2,3
TRDMT1						**rs2295809**		RFLP analysis	
TYMS	thymidylate synthetase	NT_010859.14	7	18p11.32	15.8	rs16430		APEX	2,3
TYMS						rs596909	Gly157Val	APEX	1,2
TYMS						CCGCGCCACTTGGCCTGCCTC CGTCCCG>CCGCGCCACTTCG CCTGCCTCCGTCCCG[o]		RFLP analysis	
TYMS						-97(CCGCGCCACTTGGCCT GCCTCCGTCCCG)2-4[p]		GeneScan analysis	

[a] approved symbol and name according to HUGO Gene Nomenclature Committee (http://www.genenames.org/ ; page last updated: July 16, 2007)
[b] 1: folate transport; 2: folate metabolism; 3: methylation cycle; 4: transsulfuration pathway; 5: methyl transferases; 6: polyamine biosynthesis; 7: purine and pyrimidine synthesis; 8: other
[c] size according to HapMap release 22 / phase II, April 2007 on NCBI assembly, dbSNP Build 126
[d] where available, rs number is given; otherwise, mRNA change is displayed (according to dbSNP Build 128)
[e] 1: minor allele frequency <0.02; 2: Hardy-Weinberg equilibrium P-value<0.01; 3: >25% missing genotypes
[f] gene variants associated with NTD with nominal P<0.10 are printed in bold
[g] 68 bp: CATCCAGGTGGGGTTTTGCTGGGCTTGAGCCCTGAAGCCGCGCCCTCTGCAGATCATTGGGGTGGAT
[h] for heterogeneous repeat units see Lievers et al.[18]
[i] recoded into di-allelic variant; allele 1 = 18; allele 2 = 17 or 19 or 20
[j] recoded into di-alllelic variant; allele 1 = 5 repeats; allele 2 = 8 repeats; 3 and 6 repeat alleles were very rare and excluded from analysis
[k] recoded into di-allelic variant; allele 1 = less than 12 repeats; allele 2 = 12 or more repeats
[l] recoded into di-allelic variant; allele 1 = 17 repeats; allele 2 = more or less than 17 repeats
[m] recoded into di-allelic variant; allele 1 = 8 or 9 or 10 or 11 or 12 repeats; allele 2 = 13 or 14 or 15 or 16 or 17 repeats
[n] only 2 alleles present (4 or 5 repeats) in population
[o] G>C transversion in the 3 repeat allele of the -97(CCGCGCCACTTGGCCTGCCTCCGTCCCG)2-4 in TYMS
[p] only 2 alleles present (2 or 3 repeats) in population

Table 5.2 *Association results for polymorphisms showing nominal association with spina bifida aperta (nominal Chi-square P-value <0.05) and geometric mean levels of serum and red blood cell (RBC) folate, serum vitamin B_{12} and plasma total homocysteine (tHcy) for genotype groups of selected polymorphisms in the control sample.*

			Case-control analysis				Geometric mean[b] (95% CI) metabolite concentrations in controls[c]			
Gene variant	Gene symbol	Genotype[a]	Controls, N (%)	Cases, N (%)	OR (95% CI)	Chi-square P-value	Folate (nmol/L)	RBC folate (nmol/L)	B_{12} (pmol/L)	tHcy (µmol/L)
rs1907362	*CUBN*	GG	164 (86.3)	175 (97.8)	1	0.0000577	12.5 (11.6 – 13.6)	380.1 (354.8 – 407.3)	214.4 (201.6 – 228.1)	10.1 (9.5 – 10.7)
		GA/AA	26 (13.7)	4 (2.2)	0.14 (0.05 to 0.42)		12.4 (10.6 – 14.6)	509.3 (453.8 – 571.7)*	254.3 (217.7 – 297.0)*	8.8 (7.5 – 10.4)
rs2295809	*TRDMT1*	TT/AT	123 (73.7)	152 (87.4)	1	0.0013653	12.3 (11.3 – 13.5)	376.5 (349.7 – 405.2)	220.4 (204.3 – 237.8)	10.1 (9.5 – 10.8)
		AA	44(26.4)	22 (12.6)	0.40 (0.23 to 0.71)		14.6 (12.5 – 17.1)	482.4 (413.8 – 562.2)*	221.3 (197.8 – 247.7)	9.1 (7.9 – 10.4)
rs1127717	*ALDH1L1*	AA/AG	187 (98.4)	167 (93.3)	1	0.0127207				
		GG	3 (1.6)	12 (6.7)	4.48 (1.24 to 16.15)					
rs4748353	*CUBN*	TT	167 (93.3)	165 (94.8)	1	0.0175473				
		TC	11 (6.2)	3 (1.7)	0.28 (0.08 to 1.01)					
		CC	1 (0.6)	6 (3.5)	6.07 (0.72 to 60.0)					
rs573904	*SARDH*	CC	104 (54.7)	77 (42.8)	1	0.0214443				
		CT/TT	86 (45.3)	103 (57.2)	1.62 (1.07 to 2.44)					
rs1051266	*SLC19A1*	GG/GA	149 (78.8)	157 (87.2)	1	0.0323536				
		AA	40 (21.2)	23 (12.8)	0.55 (0.31 to 0.96)					

[a] genetic model (dominant, recessive, genotype) that showed lowest P-value is displayed
[b] unadjusted geometric means
[c] only results for rs1907362 and rs2295809 are given; other depicted genetic variants did not show nominal association to any of the metabolite concentrations
* nominal P-value <0.05 for difference between genotype groups for age- and sex- adjusted values

Haplotype analysis was performed for *CUBN*, *ALDH1L1*, *SARDH*, and *BHMT*. No haplotype blocks and no strong LD between variants in these genes were found (all $r^2 < 0.8$). Sliding marker window analysis resulted in nominal associations for marker windows in *CUBN*. However, no haplotype effect on top of the single locus effect of the rs1907362 variant was present (data not shown).

The best multilocus model identified in logic regression contained the polymorphisms rs1907362 in *CUBN* and rs2295809 in *TRDMT1*. However, no indication for statistical interaction between the two SNPs was found. Logistic regression analysis confirmed the nominal significance of both effects and statistical additivity of the two variants. The two genes are direct neighbors on chromosome 10, though the two polymorphisms rs1907362 (*CUBN*) and rs2295809 (*TRDMT1*) are not highly correlated ($r^2 = 0.03$).

The analyses for the spina bifida-associated polymorphisms and levels of biochemical parameters showed nominally statistically significant associations for rs1907362 in CUBN and increased RBC folate ($p = 0.001$) and vitamin B_{12} levels ($p = 0.039$) (Table 5.2). Also, rs2295809 in *TRDMT1* was associated with increased RBC folate ($p = 0.002$).

Discussion

In this study, we identified several novel candidate genes for spina bifida aperta and could confirm some prior association findings. The two polymorphisms in *CUBN* and *TRDMT1* that were most strongly associated with spina bifida risk showed association with vitamin B_{12} and/or RBC folate levels. With the exception of *SLC19A1*, none of the genes from our top findings - *CUBN*, *TRDMT1*, *ALDH1L1* or *SARDH* - has been investigated for their association with spina bifida before, to our knowledge.

The CUBN gene encodes the intrinsic factor-cobalamin receptor (cubilin), a peripheral membrane protein that acts as a receptor for intrinsic factor-vitamin B_{12} complexes as well as many other compounds, including proteins and lipids[23]. During embryonic development, the gene mediates ligand endocytosis by absorptive epithelia at the maternal-fetal interfaces, trophectoderm and visceral endoderm, but is also expressed by developing neuroepithelial cells and the neural tube. Although intronic, a possible functional character of SNP rs1907362 (G>A), or an SNP in LD, is supported by the finding that the risk-decreasing A allele was nominally associated with increased vitamin B_{12} levels and strongly increased RBC folate levels.

TRDMT1 codes for a methyltransferase that methylates a specific RNA molecule, the aspartic acid transfer RNA (tRNAAsp)[24]. The A allele of the disease-associated rs2295809 was associated with increased RBC folate in the control population, which is in line with its risk-reducing effect observed in this study.

The *ALDH1L1* and *SARDH* genes also represent interesting new candidates for further study. A recent report on *ALDH1L1* shows midline-specific expression in the developing central nervous system of mouse embryos. The authors explicitly mention the potential of this gene for involvement in NTDs in humans[25].

Sarcosine dehydrogenase, encoded by SARDH, is involved in the breakdown of choline via betaine, catalyzing the oxidative demethylation of sarcosine to form glycine, and is important as a supplier of one-carbon units to the folate metabolism. Interestingly, intake of choline and polymorphisms in choline metabolizing genes were found to influence NTD risk in previous studies[26,27].

SLC19A1 encoding the reduced folate carrier is responsible for the uptake of dietary folate (in the form of folate monoglutamates) in the small intestine[8]. The polymorphism assessed in this study (rs1051266, 80AG) causes a change in amino acid at position 27. Several earlier studies have assessed the effect of this variant on NTD risk[28-33]. Most of them indeed showed an increased risk for carriers of the GG genotype[28-31]. These data are compatible with our findings showing a reduced NTD risk for AA carriers.

The results of our study should be viewed in the context of some strengths and limitations. Strengths include the many folate-related genes investigated, the homogeneous spina bifida aperta background of the Dutch study population, and the availability of folate, vitamin B_{12}, and tHcy measurements in healthy individuals. Limitations are the relatively small sample size (180 patients and 190 controls) and the incomplete coverage of the genes in terms of LD. Furthermore, maternal genotypes were not available, so we were not able to distinguish maternal effects from those based on the child's genotype. Our biochemical analyses indicate that a genotype has consequences on the phenotypic level in a direction that might help explain its association with NTDs, though it is clearly debatable whether such effects are comparable between adults and the developing embryo. In future studies, larger, well-characterized samples of children and mothers should be investigated for association of genes with spina bifida. In addition, information on folate supplementation in the mother should be taken into account. Preferably, such studies should be performed in a genome-wide effort, to reduce the bias introduced by current knowledge of candidate genes for spina bifida.

Acknowledgements

We would like to thank Ivon van der Linden, Pascal Groenen, Riko Klootwijk, and Huub Straatman for their valuable contributions. Funding for this study was obtained from the Dutch Prinses Beatrix Fonds, grant MAR02-0206.

References

1. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. Lancet. 1991;338:131-137.
2. Czeizel AE, Dudas I. Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation N Engl J Med. 1992;327:1832-1835.
3. Berry RJ, Li Z. Folic acid alone prevents neural tube defects: evidence from the China study. Epidemiology. 2002;13:114-116.
4. Steegers-Theunissen RP, Boers GH, Trijbels FJ, Eskes TK. Neural-tube defects and derangement of homocysteine metabolism. N Engl J Med. 1991;324:199-200.
5. Ray JG, Blom HJ. Vitamin B12 insufficiency and the risk of fetal neural tube defects. QJM. 2003;96:289-295.
6. Groenen PM, Klootwijk R, Schijvenaars MM, Straatman H, Mariman EC, Franke B, Steegers-Theunissen RP. Spina bifida and genetic factors related to myo-inositol, glucose, and zinc. Mol Genet Metab. 2004;82:154-161.
7. Blom HJ, Shaw GM, den Heijer M, Finnell RH. Neural tube defects and folate: case far from closed. Nat Rev Neurosci. 2006;7:724-731.
8. van der Linden I, Afman LA, Heil SG, Blom HJ. Genetic variation in genes of folate metabolism and neural-tube defect risk. Proc Nutr Soc. 2006;65:204-215.
9. Mariman EC, Hamel BC. Sex ratios of affected and transmitting members of multiple case families with neural tube defects. J Med Genet. 1992;29:695-698.
10. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet. 1995;346:1070-1071.
11. Klootwijk R, Groenen P, Schijvenaars M, Hol F, Hamel B, Straatman H, Steegers-Theunissen R, Mariman E, Franke B. Genetic variants in ZIC1, ZIC2, and ZIC3 are not major risk factors for neural tube defects in humans. Am J Med Genet A. 2004;124:40-47.
12. den Heijer M, Blom HJ, Gerrits WB, Rosendaal FR, Haak HL, Wijermans PW, Bos GM. Is hyperhomocysteinaemia a risk factor for recurrent venous thrombosis? Lancet. 1995;345:882-885.
13. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308-311.
14. Riva A, Kohane IS. SNPper: retrieval and analysis of human SNPs. Bioinformatics. 2002;18:1681-1685.
15. Tõnisson N, Kurg A, Kaasik K, Lõhmussaar E, Metspalu A. Unravelling genetic data by arrayed primer extension. Clin Chem Lab Med. 2000;38:165-170.
16. Cleves MA. Hardy-Weinberg equilibrium tests and allele frequency estimation. Stata Technical Bulletin. 1999;48:34-37.
17. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 1992;48:361-372.

18. Lievers KJ, Kluijtmans LA, Heil SG, Boers GH, Verhoef P, van Oppenraay-Emmerzaal D, den Heijer M. Trijbels FJ, Blom HJ. A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. Eur J Hum Genet. 2001;9:583-589.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289-300.
20. Newson R. Multiple-test procedures and smile-plots. The Stata Journal 2003;3:109-132.
21. Purcell S, Daly MJ, Sham PC. WHAP: haplotype-based association analysis. Bioinformatics. 2007;23:255-256.
22. Ruczinski I, Kooperberg C, LeBlanc ML. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. Journal of Multivariate Analysis. 2004;90:178-195.
23. Christensen EI, Birn H. Megalin and cubilin: multifunctional endocytic receptors. Nat Rev Mol Cell Biol. 2002;3:256-266.
24. Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh CL, Zhang X, Golic KG, Jacobsen SE, Bestor TH. Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science. 2006;311:395-398.
25. Anthony TE, Heintz N. The folate metabolic enzyme ALDH1L1 is restricted to the midline of the early CNS, suggesting a role in human neural tube defects. J Comp Neurol. 2007;500:368-383.
26. Shaw GM, Carmichael SL, Yang W, Selvin S, Schaffer DM. Periconceptional dietary intake of choline and betaine and neural tube defects in offspring. Am J Epidemiol. 2004;160:102-109.
27. Enaw JO, Zhu H, Yang W, Lu W, Shaw GM, Lammer EJ, Finnell RH. CHKA and PCYT1A gene polymorphisms, choline intake and spina bifida risk in a California population. BMC Med. 2006;4:36.
28. Shaw GM, Lammer EJ, Zhu H, Baker MW, Neri E, Finnell RH. Maternal periconceptional vitamin use, genetic variation of infant reduced folate carrier (A80G), and risk of spina bifida. Am J Med Genet. 2002;108:1-6.
29. De Marco P, Calevo MG, Moroni A, Merello E, Raso A, Finnell RH, Zhu H, Andreussi L, Cama A, Capra V. Reduced folate carrier polymorphism (80AG) and neural tube defects. Eur J Hum Genet. 2003;11:245-252.
30. Morin I, Devlin AM, Leclerc D, Sabbaghian N, Halsted CH, Finnell R, Rozen R. Evaluation of genetic variants in the reduced folate carrier and in glutamate carboxypeptidase II for spina bifida risk. Mol Genet Metab. 2003;79:197-200.
31. Pei L, Zhu H, Ren A, Li Z, Hao L, Finnell RH, Li Z. Reduced folate carrier gene is a risk factor for neural tube defects in a Chinese population. Birth Defects Res A Clin Mol Teratol. 2005;73:430-433.

32. Vieira AR, Murray JC, Trembath D, Orioli IM, Castilla EE, Cooper ME, Marazita ML, Lennon-Graham F, Speer M. Studies of reduced folate carrier 1 (RFC1) A80G and 5,10-methylene-tetrahydrofolate reductase (MTHFR) C677T polymorphisms with neural tube and orofacial cleft defects. Am J Med Genet A. 2005;135:220-223.
33. O'leary VB, Pangilinan F, Cox C, Parle-McDermott A, Conley M, Molloy AM, Kirke PN, Mills JL, Brody LC, Scott JM; Members of the Birth Defects Research Group. Reduced folate carrier polymorphisms and neural tube defect risk. Mol Genet Metab. 2006;87:364-369.

Sandra G. Heil
Sita H. Vermeulen
Brenda J.M. Van der Rijt-Pisa
Martin den Heijer
Henk J. Blom

Adapted version published in: Clinical Chemistry and Laboratory Medicine 2008;46:655-659

CHAPTER 6

Role for mitochondrial uncoupling protein-2 (*UCP2*) in hyperhomocysteinemia and venous thrombosis risk?

Abstract

Background:
Hyperhomocysteinemia has been associated with an increased risk of venous thrombosis, which might be mediated through an oxidative stress dependent mechanism. The function of uncoupling protein-2 (UCP2) is still under debate, but it has been suggested to play a role in reduction of mitochondrial reactive oxygen species. In the present study, we investigated whether the 45 bp deletion/insertion (del/ins) polymorphism in the *UCP2* gene is associated with elevated homocysteine levels and whether it might be associated with an increased risk of recurrent venous thrombosis (RVT).

Methods:
The 45 bp del/ins polymorphism in the *UCP2* gene was genotyped by PCR analysis in 161 RVT cases and 386 controls of Caucasian origin in which fasting-and post-load homocysteine levels were previously determined. Statistical analysis was performed to assess whether the *UCP2* 45 bp del/ins polymorphism was associated with plasma total homocysteine levels and venous thrombosis risk.

Results:
Post-load homocysteine levels were positively associated with *UCP2* 45 bp ins/ins genotype (p=0.02). None of the *UCP2* 45 bp ins/del genotypes were associated with fasting plasma homocysteine levels. The frequency of the *UCP2* 45 bp ins/ins genotype was 12.4% in RVT cases compared to 8.3% in controls, which resulted in an odds ratio of 1.8 (95% CI 1.0–3.4).

Conclusions:
The results of our study show that the common 45 bp del/ins polymorphism in the *UCP2* gene is associated with hyperhomocysteinemia, which might increase the risk of venous thrombosis. However, the mechanism is not fully understood and additional studies should be performed to confirm our findings.

Introduction

Hyperhomocysteinemia has been associated with an increased risk of deep vein thrombosis and pulmonary embolism. Both genetic and environmental factors contribute to elevated homocysteine levels. Early data from case-control studies showed a clear association between hyperhomocysteinemia and venous thrombosis[1], whereas data from prospective studies showed a much less, though still, significant increased relative risk[2]. In addition, in a recent meta-analysis it was demonstrated that the 5,10-methylenetetrahydrofolate reductase (*MTHFR*) 677 TT genotype, which is a genetic determinant of hyperhomocysteinemia, increased the risk of venous thrombosis[3]. The above-mentioned studies all provided evidence that hyperhomocysteinemia was associated with an increased risk of venous thrombosis. However, results from the first randomized controlled trials showed that homocysteine-lowering by B-vitamin supplementation does not prevent recurrent venous thrombosis[4, 5]. Currently, the question rises whether homocysteine is causally related to venous thrombosis itself or is a marker of disturbed intracellular metabolism. Studying genetic variation in candidate genes might provide an answer to the mechanism underlying the prothrombotic actions of homocysteine or related metabolites, which might involve oxidative stress, DNA hypomethylation and/ or proinflammatory effects[6].

Preliminary results of our unpublished study, in which we studied the association of homocysteine levels with single-nucleotide polymorphisms (SNP) in folate/neural tube defect related genes in 190 control individuals by a SNP array, showed a positive association of the uncoupling protein-2 (*UCP2*) gene 45 bp deletion/insertion (del/ins) polymorphism and postload homocysteine levels. This SNP array included three SNPs in the *UCP2* gene [i.e., the –866 G>A transversion (rs659366), the 45 bp del/ins polymorphism and the 544 C>T transition (rs660339)], of which only the 45 bp del/ins polymorphism was associated with post-load homocysteine levels (p=0.02) (unpublished results). These results, together with the postulated role of *UCP2* as a regulator of mitochondrial reactive oxygen species (ROS) production, were the rationale for the present study.

Mitochondrial uncoupling proteins (UCPs) are located in the mitochondrial inner membrane and the function of the best-described uncoupling protein (i.e., UCP1) involves translocation of protons from the inner membrane space to the matrix to uncouple ATP synthesis to produce heat[7]. In addition to UCP1, four other putative UCPs with unknown physiological roles have been described (UCP2–5)[7, 8]. Among these four putative UCPs, UCP2 has been the focus of much research, because it is thought to be involved in the regulation of mitochondrial ROS production rather than being a mitochondrial UCP[9]. Several variations in the *UCP2* gene have been described, of which in particular the 45 bp del/ins polymorphism in the 3'-untranslated region of exon 8 has been shown to be functional, i.e., mRNA transcribed from the insertion allele had

a shorter half-life in cultured myoblasts than mRNA transcribed from the deletion allele[10].

In this study, we investigated whether the *UCP2* 45 bp del/ins polymorphism is associated with hyperhomocysteinemia and recurrent venous thrombosis (RVT). We genotyped the *UCP2* 45 bp del/ins poly-morphism in 169 cases with RVT and 390 controls of Caucasian origin and examined its association with fasting- and post-load homocysteine levels and RVT risk.

Materials and methods

Patient material

Patients were selected from an anticoagulant clinic where they were registered as having RVT[1,11]. Patients with a history of two or more episodes of venous thrombosis (n=473) were invited to participate in the study, of which 185 agreed to take part in the study. Almost all patients received coumarin therapy. Controls were selected from the general population in The Hague. A total of 2812 individuals were approached to take part in a health survey of risk factors for cardiovascular disease and 532 agreed to take part; however, 500 individuals actually took part in the study[11]. Of these 500, only individuals that were of Caucasian origin (n=462) were included in this study. After overnight fasting, blood samples were collected before and after 6 h of an oral methionine load[11]. Plasma total homocysteine (fasting and post-methionine load) levels were measured according to standard methods as described previously[11]. In the present study, *UCP2* 45 bp del/ins genotypes were obtained for 161 RVT cases and 386 controls (these controls included the 190 controls that were initially analyzed by the SNP array and that were re-genotyped for this study). Characteristics of the initial study group are presented in Table 6.1.

Determination of the UCP2 45 bp insertion

Genomic DNA was isolated from peripheral blood lymphocytes according to standard procedures[12]. The 45 bp del/ ins polymorphism in the gene coding for *UCP2* is a 45 bp sequence that is duplicated[13]. In general, the sequence with the 45 bp fragment is referred to as deletion, whereas the sequence with a duplication of the 45 bp fragment is referred to as insertion[10]. PCR was carried out in a total volume of 50 mL on the iCycler thermocycler (Biorad, Veenendaal, The Netherlands) containing 200 nM of forward (CAGTGAGGGAAGTGGGAGGTG)[14] and reverse (GCAGGACGAAGATTCTGGCTG) primers, 200 mM of dNTPs, 1.5 mM MgCl2 (Invitrogen, Breda, The Netherlands), 5% DMSO, 1x PCR buffer (Invitrogen) and 1 U of recombinant DNA polymerase (Invitrogen). In the case of the deletion, a fragment of 453 bp was obtained, and in the presence of the insertion, a fragment of 498 bp was obtained. Fragments were separated

Table 6.1 *Baseline characteristics of study population.*

Variable*	Cases (n=185)	Controls (n=500)	p-value
Age (year)	62 (42-79)	50 (34-69)	<0.01
Sex (men)	94 (51%)	208 (42%)	0.03
Post-menopausal women	64 (35%)	138 (28%)	<0.01
Time between 1st event and study (years)	17 (range 1 to 58)	-	-
Time between last event and study (years)	7 (range 1 to 30)	-	-
Type of venous thrombosis Only pulmonary embolism (PE) Only deep vein thrombosis (DVT) Both DVT and PE	 43 73 69	-	-
Smoking (yes/no)	61 (33%)	167 (34%)	0.9
B-vitamin (yes/no)	23 (12%)	76 (15%)	0.9
Fasting homocysteine (µmol/L)	12.2 (8.3-20.8)	10.7 (6.7-15.5)	<0.01
Postload homocysteine (µmol/L)	43.6 (29.3-68.0)	37.0 (25.5-57.3)	< 0.01
Creatinine (µmol/L)	81 (61-112)	74 (55-99)	<0.01
Folate (nmol/L)	13.5 (7.7-23.5)	12.7 (7.0-23.6)	0.2
Vitamin B_{12} (µmol/L)	241 (118-466)	217 (125-389)	0.04
Vitamin B_6 (nmol/L)	25.8 (12.7-55.7)	27.6 (15.7-54.8)	0.03
Riboflavin (nmol/L)	10.2 (4.9-37.1)	9.1 (4.0-25.9)	0.07
MTHFR 677 C>T (MAF)	33.3%	29.1%	0.21

* Data are given as Medians with 10th to 90th percentiles in parenthesis for continuous variables and as numbers with percentages in parenthesis for categorical variables; MAF = minor allele frequency

by gel electrophoresis on a 2% agarose gel. In addition, four samples per genotype were sequenced on the 3130 XL DNA analyzer according to the protocol of the manufacturer (Applied Biosystems, Nieuwerkerk a/d Ijssel, The Netherlands) to validate the genotyping method.

Statistical analysis

To determine whether the *UCP2* 45 bp del/ins polymorphism was associated with elevated homocysteine levels, geometric means and confidence intervals were calculated, and a one-way analysis of variance (ANOVA) was performed. Odds ratios (ORs) and 95% confidence intervals (95% CIs) were calculated for *UCP2* 45 bp genotypes in relation to recurrent thrombosis risk. Logistic regression analysis with *UCP2* genotypes as a continuous variable was applied to test for a trend among the three genotypes in relation to RVT risk. A $p<0.05$ was considered statistically significant; all p-values were two-tailed.

Table 6.2 *Fasting- and postload total homocysteine levels for different UCP2 genotypes.*

	UCP2 45bp del/ins			ANOVA p
	DEL/DEL Mean[a] (95%CI)[N]	DEL/INS Mean[a] (95%CI)[N]	INS/INS Mean[a] (95%CI)[N]	
Controls (n=386)				
Fasting tHcy (µmol/L)	10.3 (9.8-10.8)[207]	10.9 (10.2-11.7)[147]	10.5 (9.4-11.7)[32]	0.38
Post-load tHcy (µmol/L)	37.1 (35.5-38.8)[206]	40.2 (38.2-42.3)[145,b]	41.9 (36.7-47.8)[32,b]	0.02
Cases (n= 161)				
Fasting tHcy (µmol/L)	12.4 (11.4-13.5)[71]	13.4 (12.1-14.9)[70]	12.3 (10.3-14.7)[20]	0.42
Post-load tHcy (µmol/L)	42.4 (39.5-45.6)[71]	46.8 (43.3-50.6)[70,c]	40.6 (34.6-47.7)[20]	0.09

[a] Geometric mean

[b] P< 0.05 compared to *UCP2* 45bp del/del genotype

[c] P=0.07 compared to *UCP2* 45bp del/del genotype

tHcy = plasma total homocysteine

Table 6.3 *Genotype frequencies of UCP2 45bp del/ins polymorphism in RVT cases and controls.*

UCP2 45bp del/ins	Controls N (%)	Cases N (%)	OR [95%CI]
del/del	207 (53.6%)	71 (44.1%)	1.0[a]
del/ins	147 (38.1%)	70 (43.5%)	1.4 [0.9-2.1]
ins/ins	32 (8.3%)	20 (12.4%)	1.8 [1.0-3.4][b]

[a] Reference category

[b] P_{trend} = 0.03

Results

We examined whether plasma total homocysteine (fasting and/or post-load) levels were different between cases and controls with different *UCP2* 45 bp del/ins genotypes. Presently, DNA was available from 169 of the 185 RVT cases and 390 of the 462 Caucasian controls. Genotype data were obtained from 161 RVT cases and 386 controls of which 51% and 41% were male, respectively (p=0.04). Mean age of cases and controls was 62.3±14.1 years and 51.0±13.4 years, respectively (p=0.01). Mean plasma fasting homocysteine levels were 12.9±6.5 µmol/L for cases and 10.5±4.7 µmol/L for controls (p=0.01). Post-load homocysteine levels were 44.0±15.5 µmol/L for cases and 38.6±14.5 µmol/L for controls (p=0.01).

UCP2 45 bp del/ins allele frequencies did not differ from that expected under Hardy-Weinberg equilibrium (p=0.09). Fasting homocysteine levels were not different between RVT cases and controls with different *UCP2* genotypes (p=0.42 and p=0.38, respectively; Table 6.2). Post-load homocysteine levels were significantly elevated among controls with the *UCP2* 45 bp del/ins and ins/ins genotypes compared to those with the *UCP2* del/del genotype (p=0.02, Table 6.2). In RVT cases, a slight association was observed between the *UCP2* 45 bp del/ins genotype with post-load homocysteine levels compared to the *UCP2* 45 bp del/del genotype, although this was not significant (p=0.07, Table 6.2).

In addition, we examined whether the *UCP2* 45 bp ins/ins genotype occurred more often in RVT cases than in controls. *UCP2* 45 bp del/ins genotyping was performed as described in genomic DNA of 169 RVT cases, of which a reliable genotype could be obtained from 161 cases. The *UCP2* 45 bp ins/ins genotype was present in 12.4% of the RVT cases and 8.3% of the controls, which led to an OR of 1.8 (95% CI 1.0–3.4) (Table 6.3). A positive trend was observed for the *UCP2* 45 bp insertion allele in relation to RVT risk (p_{trend}=0.03). Correction for age and sex differences between RVT cases and controls did not change the outcome of this analysis (data not shown).

Discussion

In this study, we tested the hypothesis that the common 45 bp del/ins polymorphism in the *UCP2* gene is associated with hyperhomocysteinemia and RVT risk. We found a positive association of the *UCP2* 45 bp ins/ins genotype with post-load homocysteine levels and showed that this genotype was more frequently present among RVT cases than controls.

The role of UCP2 is still under debate but one of its suggested roles is protection against mitochondrial ROS production[8, 15, 16]. In a previous unpublished study, we found evidence for a positive association between the 45 bp insertion in the *UCP2* gene and hyperhomocysteinemia. As hyperhomocysteinemia is suggested to be involved in oxidative stress, we investigated the role of the *UCP2* 45 bp del/ins polymorphism in relation to fasting-and post-load total homocysteine levels. In controls, we found a positive association between the *UCP2* 45 bp del/ins and ins/ins genotypes with post-load homocysteine levels in comparison with the *UCP2* 45 bp del/del genotype. In cases, we observed a positive association with post-load homocysteine levels in relation to the *UCP2* 45 bp del/ins genotype but not to the *UCP2* ins/ins genotype, which might be explained by the small number of cases with the *UCP2* ins/ins genotype (n=10, Table 6.2). This is the first study that reports an association between *UCP2* and hyperhomocysteinemia and therefore more studies are warranted to confirm our results.

As hyperhomocysteinemia is associated with increased RVT risk, we hypothesized that if the 45 bp del/ins polymorphism is associated with hyperhomocysteinemia this also increases RVT risk. The *UCP2* ins/ins genotype was more frequently present among RVT cases than in controls, which led to a 1.8-fold almost significant risk compared to the *UCP2* del/del genotype. However, to confirm our findings, the results of this study need to be confirmed in a second, preferably larger case-control study, in which also gene-gene (e.g., *MTHFR*) and gene-environmental (e.g., folate, ROS) interactions can be studied.

Recent studies suggest that hyperhomocysteinemia is associated with mitochondrial dysfunction[17]. Homocysteine infusion in rats was shown to inhibit complex I activity of mitochondrial oxidative phosphorylation, which led to enhanced ROS production[18]. In addition, in cultured endothelial cells it was shown that homocysteine infusion led to the translocation of proteins to the mitochondria, which might be caused by an increased intramitochondrial oxidative burst[19, 20]. The role of *UCP2* in relation to ROS production is still under debate, but it is expected to be involved in mitochondrial function some way or another[8]. Our results with *UCP2* in humans are in line with the findings from in vitro studies that hyperhomocysteinemia is associated with altered mitochondrial functioning. However, results from homocysteine infusion studies suggest that homocysteine is causally related to mitochondrial dysfunction, whereas our results suggest the opposite (i.e., that homocysteine is a reflection of mitochondrial dysfunction). Future studies should investigate this more closely.

This is the first study that examined a possible association between *UCP2*, hyperhomocysteinemia and RVT risk. The results of our study show that hyperhomocysteinemia is associated the 45 bp del/ins polymorphism in the *UCP2* gene, which might increase RVT risk.

Acknowledgements

The authors would like to thank Mascha Schijvenaars for technical assistance. This study was supported by grants from the Dutch Kidney Foundation (C042083), the Netherlands Heart Foundation (2002B68) and the Netherlands Organization for Scientific Research (VENI grant NWO awarded to MdH).

References

1. den Heijer M, Blom HJ, Gerrits WB, Rosendaal FR, Haak HL, Wijermans PW, Bos GM. Is hyperhomocysteinaemia a risk factor for recurrent venous thrombosis? Lancet. 1995;345:882–885.

2. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. Br Med J. 2002;325:1202.
3. den Heijer M, Lewington S, Clarke R. Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost. 2005;3:292–299.
4. den Heijer M, Willems HP, Blom HJ, Gerrits WB, Cattaneo M, Eichinger S, Rosendaal FR, Bos GM. Homocysteine lowering by B vitamins and the secondary prevention of deep-vein thrombosis and pulmonary embolism. A randomized, placebo-controlled, double blind trial. Blood. 2007;109:139–144.
5. Ray JG, Kearon C, Yi Q, Sheridan P, Lonn E. Homocysteine-lowering therapy and risk for venous thromboembolism: a randomized trial. Ann Intern Med. 2007;146:761–767.
6. Lentz SR. Mechanisms of homocysteine-induced atherothrombosis. J Thromb Haemost. 2005;3:1646–1654.
7. Mattiasson G, Sullivan PG. The emerging functions of UCP2 in health, disease, and therapeutics. Antioxid Redox Signal. 2006;8:1–38.
8. Nedergaard J, Cannon B. The 'novel' 'uncoupling' proteins UCP2 and UCP3: what do they really do? Pros and cons for suggested functions. Exp Physiol. 2003;88:65–84.
9. Krauss S, Zhang CY, Lowell BB. The mitochondrial uncoupling-protein homologues. Nat Rev Mol Cell Biol. 2005;6:248–261.
10. Esterbauer H, Schneitler C, Oberkofler H, Ebenbichler C, Paulweber B, Sandhofer F, Ladurner G, Hell E, Strosberg AD, Patsch JR, Krempler F, Patsch W. A common polymorphism in the promoter of UCP2 is associated with decreased risk of obesity in middle-aged humans. Nat Genet. 2001;28:178–183.
11. Keijzer MB, den Heijer M, Blom HJ, Bos GM, Willems HP, Gerrits WB, Rosendaal FR. Interaction between hyperhomocysteinemia, mutated methylenetetrahydrofolatereductase (MTHFR) and inherited thrombophilic factors in recurrent venous thrombosis. Thromb Haemost. 2002;88:723–728.
12. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16:1215.
13. Tu N, Chen H, Winnikes U, Reinert I, Marmann G, Pirke KM, Lentes KU. Structural organization and mutational analysis of the human uncoupling protein-2 (hUCP2) gene. Life Sci. 1999;64:L41–50.
14. Walder K, Norman RA, Hanson RL, Schrauwen P, Neverova M, Jenkinson CP, Easlick J, Warden CH, Pecqueur C, Raimbault S, Ricquier D, Silver MH, Shuldiner AR, Solanes G, Lowell BB, Chung WK, Leibel RL, Pratley R, Ravussin E. Association between uncoupling protein polymorphisms (UCP2-UCP3) and energy metabolism/obesity in Pima indians. Hum Mol Genet. 1998;7:1431–1435.

15. Arsenijevic D, Onuma H, Pecqueur C, Raimbault S, Manning BS, Miroux B, Couplan E, Alves-Guerra MC, Goubern M, Surwit R, Bouillaud F, Richard D, Collins S, Ricquier D. Disruption of the uncoupling protein-2 gene in mice reveals a role in immunity and reactive oxygen species production. Nat Genet. 2000;26:435–439.
16. Cannon B, Shabalina IG, Kramarova TV, Petrovic N, Nedergaard J. Uncoupling proteins: a role in protection against reactive oxygen species – or not? Biochim Bio-phys Acta. 2006;1757:449–458.
17. Tyagi N, Moshal KS, Ovechkin AV, Rodriguez W, Steed M, Henderson B, Roberts AM, Joshua IG, Tyagi SC. Mitochondrial mechanism of oxidative stress and systemic hypertension in hyperhomocysteinemia. J Cell Biochem. 2005;96:665–671.
18. Folbergrová J, Jesina P, Drahota Z, Lisý V, Haugvicová R, Vojtísková A, Houstěk J. Mitochondrial complex I inhibition in cerebral cortex of immature rats following homocysteic acid-induced seizures. Exp Neurol. 2007;204:597–609.
19. Moshal KS, Singh M, Sen U, Rosenberger DS, Henderson B, Tyagi N, Zhang H, Tyagi SC. Homocysteine-mediated activation and mitochondrial translocation of calpain regulates MMP-9 in MVEC. Am J Physiol Heart Circ Physiol. 2006; 291:H2825–2835.
20. Tyagi N, Moshal KS, Sen U, Lominadze D, Ovechkin AV, Tyagi SC. Ciglitazone ameliorates homocysteine-mediated mitochondrial translocation and matrix metalloproteinase-9 activation in endothelial cells by inducing peroxisome proliferator activated receptor-gamma activity. Cell Mol Biol. (Noisy-le-grand) 2006;52:21–27.

PART 2

Genetic epidemiological designs and analyses

Sita H Vermeulen
Min Shi
Clarice R Weinberg
David M Umbach

Published in: Genetic Epidemiology 2009;33:136-144

CHAPTER 7

A Hybrid Design: Case-Parent Triads Supplemented by Control-Mother Dyads

Abstract

Hybrid designs arose from an effort to combine the benefits of family-based and population-based study designs. A recently proposed hybrid approach augments case-parent triads with population-based control-parent triads, genotyping everyone except the control offspring. Including parents of controls substantially improves statistical efficiency for testing and estimating both offspring and maternal genetic relative risk parameters relative to using case-parent triads alone. Moreover, it allows testing of required assumptions. Nevertheless, control fathers can be hard to recruit, whereas control offspring and their mothers may be readily available. Consequently, we propose an alternative hybrid design where offspring-mother pairs, instead of parents, serve as population-based controls. We compare the power of our proposed method with several competitors and show that it performs well in various scenarios, though it is slightly less powerful than the hybrid design that uses control parents. We describe approaches for checking whether population stratification will bias inferences that use controls and whether the mating-symmetry assumption holds. Surprisingly, if mating symmetry is violated, even though mating-type parameters cannot be directly estimated using control-mother dyads alone, and maternal effects cannot be estimated using case-parent triads alone, combining both sources of data allows estimation of all the parameters. This hybrid design can also be used to study environmental influences on disease risk and gene-by-environment interactions.

Introduction

Genetic factors may increase risk for congenital disorders via direct effects of the inherited genotype of the offspring or via the genotype of the mother. Maternal effects can arise because the genetic variants carried by the mother influence the prenatal environment in which the fetus develops, thereby increasing disease risk. Under such a mechanism, the genotype distribution in the case mothers is different from that in the control population, whereas transmission of the genetic variant to the offspring conforms to Mendelian expectation. Disentangling offspring from maternal genetic effects can increase insight into the etiology of disorders with early onset in life, such as congenital malformations and childhood cancers.

The case-parents design, comprising affected offspring and their two parents, permits testing and estimation of offspring- and maternally mediated genetic effects[1,2]. Offspring effects are evaluated by the apparent over-transmission (compared to the Mendelian expectation of 0.5) of deleterious alleles from heterozygous parents to affected offspring. Maternal genotype effects are assessed through deviations from genotype mating symmetry in the case-parents data under the assumption of mating symmetry in the source population. That is, a deleterious allele that acts via the mother will be more prevalent in mothers than in fathers of affected offspring. The key mating-symmetry assumption cannot, however, be evaluated using case-parent triads alone.

Another design that can be used to disentangle offspring and maternal genetic effects ascertains affected children and their mothers as well as a random sample of unaffected children and their mothers. In such a case-mother/control-mother design, the offspring-mother pairs are the units for analysis. Both offspring and maternal effects are assessed through case-control comparisons; no mating-symmetry assumption is required. Unlike the case-parents design, however, this design is vulnerable to bias due to the existence of genetically distinct subpopulations. Weinberg and Umbach[3] describe the strengths and weaknesses of both the population-based case-mother/control-mother and the family-based case-parents design.

To bring together the advantages of both family-based and population-based approaches, Nagelkerke et al.[4], Epstein et al.[5] and Weinberg and Umbach[3] have proposed various hybrid designs that combine both kinds of data. One hybrid approach[3] enrolls case-parent and control-parent triads and genotypes case-parent triads but only the parents of control offspring (case-parent triad/control parents design). This hybrid design greatly improves power for the evaluation of offspring-and maternally mediated genetic effects. It allows testing for bias from population stratification; and, if bias is detected, confounding is avoided by using only the case-parent triads. In addition, the assumption of parental mating symmetry in the population at large can be tested, and maternal genetic effects can be validly estimated even if mating symmetry is rejected[3].

The case-parent triad/control-parents hybrid design calls for both control parents to be genotyped. Fathers, however, are often hard to recruit while mothers and unaffected offspring may already be available. This paper presents an alternative hybrid design where control-mother dyads replace parents of controls. Exposure information would still be collected for both the case offspring and the unrelated control offspring. In the hybrid design that genotypes control parents, one can directly estimate the mating-type frequencies in the source population. In the proposed design that genotypes control-mother dyads instead, the information on those frequencies is indirect because the genotype of the control child acts as a surrogate for the father's genotype. This feature introduces some challenges. We demonstrate that, even in scenarios with missing genotype data, this alternative hybrid design has a greater power than the case-parents or the case-mother/control-mother designs but has slightly reduced power compared to the hybrid design that uses control parents. We also show that this alternative hybrid design retains the ability to test for bias due to population stratification and, thereby, to examine whether case-parents and control data can be safely combined. Finally, we describe procedures for checking the assumption of mating symmetry and for making valid inference on maternal effects even when that assumption fails.

Analysis of case-parent triad/control-mother dyad data

Our alternative hybrid approach starts with a random sample of affected individuals and a random sample of unaffected individuals. Cases and their parents are enrolled and genotyped but only mothers of controls and the controls themselves are genotyped. We are interested in testing for association between disease risk and the offspring and maternal genotypes at a di-allelic autosomal locus. We assume that the disease is rare and that Mendelian transmission probabilities hold for that locus in the underlying population, and hence among controls. Validly combining information from case and control families requires an assumption that any population structure is benign with respect to bias, an assumption that can be probed with the data at hand. Neither Hardy-Weinberg equilibrium (HWE) nor random mating is required for validity. Let p denote the frequency of the minor or ''variant'' allele. Which allele is designated as the ''variant'' has no effect on estimation or testing beyond the mathematical inversions. Let *M*, *F* and *C* represent the number of variant alleles (0,1,2) carried by the mother, father and child, respectively. *D* is an indicator variable for disease status, which is 1 for case families and 0 for control families.

Following Schaid and Sommer[6], we define nine different mating types based on the number of variant copies carried by the mother and the father. These mating types along with their possible offspring genotypes lead to 15 possible (*M,F,C*) categories,

Table 7.1 *Expected counts of case-parent triads and control-mother dyads under mating asymmetry or mating symmetry.*

			mating asymmetry		mating symmetry	
MF[a]	C[b]	affected	mating types	expected count[c]	mating types	expected count[c]
00	0	yes	00	$B\mu_{00}$	00	$B\mu_{00}$
01	0	yes	01	$B\mu_{01}/2$	01/10	$B\mu_{01}/4$
01	1	yes	01	$B\mu_{01}R_1/2$	01/10	$B\mu_{01}R_1/4$
10	0	yes	10	$B\mu_{10}S_1/2$	01/10	$B\mu_{01}S_1/4$
10	1	yes	10	$B\mu_{10}R_1S_1/2$	01/10	$B\mu_{01}R_1S_1/4$
02	1	yes	02	$B\mu_{02}R_1$	02/20	$B\mu_{02}R_1/2$
20	1	yes	20	$B\mu_{20}R_1S_2$	02/20	$B\mu_{02}R_1S_2/2$
11	0	yes	11	$B\mu_{11}S_1/4$	11	$B\mu_{11}S_1/4$
11	1	yes	11	$B\mu_{11}R_1S_1/2$	11	$B\mu_{11}R_1S_1/2$
11	2	yes	11	$B\mu_{11}R_2S_1/4$	11	$B\mu_{11}R_2S_1/4$
12	1	yes	12	$B\mu_{12}R_1S_1/2$	12/21	$B\mu_{12}R_1S_1/4$
12	2	yes	12	$B\mu_{12}R_2S_1/2$	12/21	$B\mu_{12}R_2S_1/4$
21	1	yes	21	$B\mu_{21}R_1S_2/2$	12/21	$B\mu_{12}R_1S_2/4$
21	2	yes	21	$B\mu_{21}R_2S_2/2$	12/21	$B\mu_{12}R_2S_2/4$
22	2	yes	22	$B\mu_{22}R_2S_2$	22	$B\mu_{22}R_2S_2$
0-[d]	0	no	00 or 01	$\mu_{00}+\mu_{01}/2$	00 or 01/10	$\mu_{00}+\mu_{01}/4$
0-	1	no	01 or 02	$\mu_{01}/2+\mu_{02}$	01/10 or 02/20	$\mu_{01}/4+\mu_{02}/2$
1-	0	no	10 or 11	$\mu_{10}/2+\mu_{11}/4$[e]	01/10 or 11	$\mu_{01}/4+\mu_{11}/4$[e]
1-	1	no	10 or 11 or 12	$\mu_{10}/2+\mu_{11}/2+\mu_{12}/2$[e]	01/10 or 11 or 12/21	$\mu_{01}/4+\mu_{11}/2+\mu_{12}/4$[e]
1-	2	no	11 or 12	$\mu_{11}/4+\mu_{12}/2$[e]	11 or 12/21	$\mu_{11}/4+\mu_{12}/4$[e]
2-	1	no	20 or 21	$\mu_{20}+\mu_{21}/2$	02/20 or 12/21	$\mu_{02}/2+\mu_{12}/4$
2-	2	no	21 or 22	$\mu_{21}/2+\mu_{22}$	12/21 or 22	$\mu_{12}/4+\mu_{22}$

[a] Copies of variant allele carried by the mother and the father.

[b] Copies of variant allele carried by the child.

[c] "B" is a normalizing factor whose inclusion ensures that the total fitted count for case triads matches the corresponding total observed count and, similarly, for control dyads. Under mating symmetry, stratification parameter μ_{mf} includes the stratum with *m* and *f* reversed.

[d] "-"is used to denote a missing paternal genotype.

[e] When *M*=1, the expected counts in cells where *C*=0 and *C*=2 sum to the expected count in the cell where *C*=1.

and, consequently, one can imagine two 15-cell multinomial distributions of offspring and parental genotypes, one for control triads and one for case triads. In hybrid designs, typically the full (*M,F,C*) data are recorded for case families, whereas only partial data are collected from control families; here, only (*M,C*) data. Initially, we assume that

any genotyping called for by the design is complete, but missing-data methods can be employed if some genotypes are missing[7].

For control families, expected counts in the 15-cell multinomial can be modeled using Mendelian transmission probabilities and mating-type parameters (μ_{mf}), which are proportional to the frequencies of mother-father pairs with $M = m$ and $F = f$ in the source population. The expected counts for control-mother dyads (the last 7 lines of Table 7.1) arise from the 15-cell multinomial by summing counts across the genotypes of possible fathers. The distribution of control-mother dyads has two noteworthy features: First, the (*M,C*) cells (0,2) and (2,0) are not possible, leaving only seven cells with non-zero expected counts. Second, and less obviously, the following relationship is a consequence of Mendelian transmission alone: when $M = 1$, the expected count for $C = 1$ is the sum of the expected counts for $C = 0$ and 2. This constraint reduces the available degrees of freedom (dfs) contributed by the seven control-dyad cells from six to five.

For case families, expected counts in the 15-cell multinomial involve not only mating-type parameters and Mendelian probabilities but also four genetic relative risk parameters. We denote these relative risk parameters as follows: R_1 (R_2) is the relative risk for offspring carrying one (respectively, two) copies of the variant compared to offspring carrying none; S_1 (S_2) is the relative risk for offspring whose mother carries one (respectively, two) copies of the variant allele compared to offspring whose mother carries none. Combining the 15 cells for case-parent triads with the 7 cells for control-mother dyads yields a 22-cell multinomial for the proposed design (Table 7.1).

For the case-parents design and for the hybrid design that uses parents of controls, the multinomial expected cell counts are all products of parameters and can be fitted using log-linear Poisson regression. Because the expected counts for control-mother dyads involve sums of parameters, however, the expected counts in Table 7.1 for the 22-cell multinomial are not themselves log-linear. A straight-forward way to proceed with fitting either model is to regard the 22-cell multinomial as a version of the full 30-cell multinomial (15 cells each for case families and control families), that is missing genotype data for fathers of controls by design. Thus, one can use missing-data methods like the Expectation-Maximization (EM) algorithm[8] in conjunction with a log-linear model for the 30-cell multinomial. Use of the EM algorithm also allows inclusion of any case-parent triads or control-mother dyads with missing genotypes that may arise through genotyping failure or incomplete ascertainment. For valid analysis, one must be able to assume that these genotype data are missing at random conditional on disease status and the observed genotypes. When control fathers are missing only by design, this assumption is satisfied without doubt because all of them are missing.

Assuming a multiplicative model for risk and no bias from population structure, a log-linear model for the full 30-cell multinomial (corresponding to Table 7.1, column 5) would be:

$$\ln[E(\text{count} \mid M = m, F = f, C = c, D = d)] = \ln(\mu_{mf}) + ß_1 dI_{(c=1)} + ß_2 dI_{(c=2)} + \alpha_1 dI_{(m=1)} + \alpha_2 dI_{(m=2)} + \gamma d + \ln(\text{Off}_{mfc}) \quad (1)$$

$I_{()}$ is an indicator function which takes a value of 1 if the parenthetical condition is met and 0 otherwise. ß1 and ß2 denote the natural logarithms of the offspring genetic relative risks, R_1 and R_2, respectively, and α_1 and α_2 denote the natural logarithms of maternal genetic relative risks, S_1 and S_2, respectively. γ corresponds to the natural logarithm of the normalizing factor B. The offset, denoted by Off_{mfc}, is the constant multiplier (1, ½ or ¼) given in Table 7.1, column 5. In this model, the nine mating-type parameters, μ_{mf}, are common to both cases and controls and can be interpreted as proportional to the mating-type frequencies in the source population.

This model is fundamental to the analysis of data from our proposed design. Modifications of the model by omitting or including certain additional terms allow one to construct likelihood ratio tests (LRTs) of hypotheses about the genetic relative risk parameters or to test the assumptions about bias from population stratification or about mating symmetry. In addition, models addressing maternal-fetal incompatibility can be constructed by including two additional relative risk parameters[9]. All tests of assumptions that we subsequently develop under the four risk parameters in model (1) can be developed and applied without difficulty when these two additional risk parameters are present.

LRTs in this missing-data situation must be based on the observed-data likelihood, not the pseudo-complete-data likelihood. We have used the program LEM[10] in our subsequent analyses. LEM was designed to fit log-linear models with missing data via the EM algorithm. Consequently, dealing with missing fathers among controls or with other patterns of missing data does not require special programming. The program allows incorporation of the Mendelian constraints and returns valid test statistics as well as valid estimates for risk parameters and their standard errors. Examples of the LEM scripts that we used are available at http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads.

Association of offspring or maternal genotypes with risk

An LRT statistic for any particular subset of the four relative risk parameters can be calculated by computing twice the change in maximized observed-data log like-lihood between a version of model (1) that estimates all parameters and a version that fixes the subset to be tested at their null values (most often, zero). For calculating a P value, this LRT statistic is referred to a X^2 distribution whose dfs equal the number of parameters in the subset. Model (1) can be modified, if desired, to accommodate specific modes of inheritance like recessive, dominant or log-additive.

To this point, we focused on model (1) with nine mating-type parameters (Table 7.1, column 5). We call this model the mating-asymmetry model because it places no constraints on the mating-type parameters. Mating symmetry is frequently assumed[2,6] and is required for inference on maternal genetic effects with a case-parents design. Mating symmetry means that the probability of parents with $M = m$ and $F = f$ is the same as the probability of parents with $M = f$ and $F = m$ in the source population. In terms of the μ_{mf} parameters, mating symmetry implies three constraints, namely: $\mu_{01} = \mu_{10}$, $\mu_{02} = \mu_{20}$, and $\mu_{12} = \mu_{21}$, in effect reducing nine mating types to six (Table 7.1, column 6). The mating-symmetry model can also be described algebraically using model (1) with the understanding that the mating-type parameters and offsets are changed to those of Table 7.1, column 7. Any tests that can be carried out on the relative risk parameters under the original nine-mating-type model (mating asymmetry) can also be carried out under the six-mating-type model (mating symmetry). When mating symmetry holds in the general population, tests under the mating-symmetry model will be more powerful than tests under the mating-asymmetry model because the former model involves fewer parameters. Using either nine or six mating types, model (1) employs the same mating-type parameters for both cases and controls and, thereby, relies on an assumed absence of bias from population structure, as does any case-control study.

Checking assumptions: absence of bias from population structure

For assessing the contribution of variant alleles to risk, the key to the power gains possible with this hybrid design is that control-mother dyads contribute to the estimation of the mating-type parameters. The two sets of mating-type parameters will truly be equal only if there is no bias due to population structure. The existence of such a bias requires not only that genetic population structure be present but also that the risk in noncarriers be correlated with allele frequency across subpopulations. Whenever this bias is present, model (1) is not valid for combining data from case and control families.

A test for bias from population structure is simply a test of whether model (1) with common mating-type parameters for cases and controls fits the data as well as an extended model with distinct mating-type parameters for cases and controls. Such a test could be conducted under either mating asymmetry or mating symmetry; and, in principle, the extended model would involve eight or five additional parameters, respectively. The ability to estimate separate mating-type parameters for cases and controls is constrained, however, by the use of control-mother dyads and, under mating asymmetry, by the complete aliasing of certain mating types with maternal effects. Consequently, the extended model has only three additional dfs under mating asymmetry and only four under mating symmetry. An alternative approach that

will sometimes be more powerful under mating symmetry than the four-df LRT is to employ a test for a particular linear trend in the disparity between case and control mating-type parameters. This one-df LRT compares model (1) to a model that includes the single additional term $\theta(M+F)d$ where θ, the unknown trend slope, is zero in the absence of bias from population structure. A monotone trend would arise, for example, if a population consisted of two distinct subpopulations, each in HWE at the single nucleotide polymorphism in question, and their respective baseline risks and allele frequencies were correlated. If under either of the testing strategies the fit is statistically significantly improved by the inclusion of the separate mating-type parameter(s), case and control data cannot be validly combined, and the estimation of relative risk parameters should be based on the case-parent triads only.

Checking assumptions: mating symmetry

Unlike the case-parents design where mating symmetry is required for assessing maternal effects but cannot be checked, the proposed hybrid design allows one to assess the assumption and to impose it or not accordingly. Enforcing mating symmetry when it holds will enhance power, but enforcing it when it fails could bias estimates of genetic effects, particularly maternal effects. If bias from population structure is absent, the three constraints on the nine μ_{mf} implied by mating symmetry can be tested in full. The appropriate three-df LRT is constructed by comparing model (1) under mating asymmetry (nine mating-type parameters) to model (1) under mating symmetry (six mating-type parameters). If mating symmetry is rejected, then the mating-asymmetry version of model (1) is used to test and estimate relative risk parameters.

Ideally, one might like to examine mating symmetry in the source population without concern for population structure. Unfortunately, the control-mother dyads do not supply enough information about the nine mating-type parameters under asymmetry to fully probe the three mating-symmetry constraints. A partial examination of these constraints is possible, however. Under mating symmetry, the sum of expected counts in (M,C) cells (0,1) and (1,2) equals the sum in cells (1,0) and (2,1) (Table 7.1, column 7). Re-expressed in terms of the μ_{mf}, this constraint becomes $\frac{1}{2}\mu_{01} + \frac{1}{2}\mu_{12} + \mu_{02} = \frac{1}{2}\mu_{10} + \frac{1}{2}\mu_{21} + \mu_{20}$. Thus, a one df LRT can be constructed by comparing a model in which the two sums are constrained to be equal to a model without that constraint. This test probes mating symmetry only in part because the two sums can be the same even when mating symmetry fails. This test, though limited, would be important in settings where the presence of bias from population structure forced one to rely on the case-parents component of the hybrid design for inference about risk parameters, but a check on the assumption required for examining maternal effects was desired. Interestingly, this test is closely related to the 1-TDT[11]. When applied to case-mother dyads,

the 1-TDT tests for offspring genetic effects assuming both mating symmetry and the absence of maternal genetic effects; when applied to control-mother dyads instead, the 1-TDT is asymptotically equivalent to our one-df LRT for mating symmetry.

Power comparisons: relative risk parameters

We compared the power of several study designs for detecting offspring and maternal genetic associations with disease: the family-based case-parents design, the population-based case-mother/control-mother design and two hybrid designs (case-parent triads/control parents and case-parent triads/control-mother dyads). For all these designs, we assessed power for a four-df LRT of the null hypothesis that $R_1 = R_2 = S_1 = S_2 = 1$ under several alternative risk scenarios. For the two hybrid designs, we investigated the tests under both the mating-symmetry and mating-asymmetry models. We used traditional logistic regression for the case-mother/control-mother design; we used log-linear Poisson regression for the other designs.

Our power calculations are based on the noncentrality parameter for a four-df X^2 LRT. We calculated the noncentrality parameter as the LRT statistic constructed by treating expected counts under the specified alternative hypothesis as data[12]. Values of the noncentrality parameter can be translated to power values using the noncentral X^2 distribution with four dfs. To calculate expected counts, we considered populations with allele frequencies ranging over the interval from zero to one where parental mating-type frequencies obeyed HWE and, consequently, exhibited mating symmetry. (Note HWE is simply a convenience here; it is not needed for the validity of our analyses.) We based all our calculations on 150 case families and 150 control families (the case-parents design used only the 150 case families). We plotted the noncentrality parameters as a function of allele prevalence and included horizontal reference lines corresponding to specific levels of power for a four-df LRT at a level 0.05. When the noncentrality parameter exceeds a given reference line, the LRT's power exceeds the specified power. Transformation of the noncentrality parameter to a different planned sample size with the same case:control ratio (say, for K cases) is accomplished by multiplying the noncentrality parameter values from our figure by K/150. For any a level, the ratio of two noncentrality parameters corresponds to the relative efficiencies of the corresponding two designs, i.e., the ratio of the sample sizes needed to achieve the same power.

We considered four distinct risk scenarios. The first scenario included a gene-dose effect of the fetal genotype but no effect of the maternal genotype, specifically, R_1, R_2, S_1, S_2 were 2, 3, 1, 1, respectively. The second scenario included a gene-dose effect of the maternal genotype and no effect of the fetal genotype, specifically, R_1, R_2, S_1, S_2 were 1, 1, 2, 3, respectively. The third involved recessive effects of both the fetal

and maternal genotypes (R_1, R_2, S_1, S_2 were 1, 2, 1, 3, respectively). The final scenario included a recessive effect of the fetal genotype and a dominant effect of the maternal genotype (R_1, R_2, S_1, S_2 were 1, 3, 2, 2, respectively). In all four scenarios, either hybrid design exhibited better power across the range of allele frequencies than both the case-parent triads and the case-mother/ control-mother design (Fig. 7.1). In addition, the hybrid using parents of controls always performed somewhat better than the hybrid using control-mother dyads, reflecting that the former design provides more mating-type information from control families. Analyses of the hybrid designs that enforced mating symmetry generally provided more power than those that did not, but the difference in power depended on the scenario. In the scenario with maternal but without offspring effects (Fig. 7.1, panel b), the power of hybrid designs when mating symmetry was not enforced was the same or close to that of the case-mother/control-mother design. The relative efficiency of the case-parents design compared to the case-mother/control-mother design depended on the particular scenario.

We also examined the same set of four scenarios when each individual genotype in every triad or dyad independently had a 20% chance of being missing (Fig. 7.1, right column). Thus, for example, a triad could be complete, or have exactly one or two or all three of its three genotypes missing. The missing genotypes could be the mother, father or offspring. Our analyses used all families except those where every individual had a missing genotype. For these analyses, we used the EM algorithm to recover information from families with only partial genotype data. For the case-mother/control-mother design that used logistic regression, invoking the EM algorithm entails exploiting the well-known equivalence between logistic models with discrete covariates and log-linear models for multi-way tables[12]. Although missing data lowered all the curves compared to their complete-data counterparts (Fig. 7.1), the relative efficiencies among curves representing the various designs and models were much the same. The one exception was the case-parents design; in the scenarios with offspring effects, its efficiency was more reduced by missing data than was that of the other designs.

Power for checking assumptions

We examined the power of our tests for bias due to population structure for a configuration where mating-type parameters differed between cases and controls. We created this configuration by mixing two subpopulations, each in HWE, under a null relative risk scenario[3]: the first had allele frequency 0.1 and baseline disease risk (risk among those with 0 copies of the variant) 0.001; the second had allele frequency 0.5 and baseline disease risk 0.003. This mixture induced a trend in the discrepancy between the separate case and the control mating-type parameters.

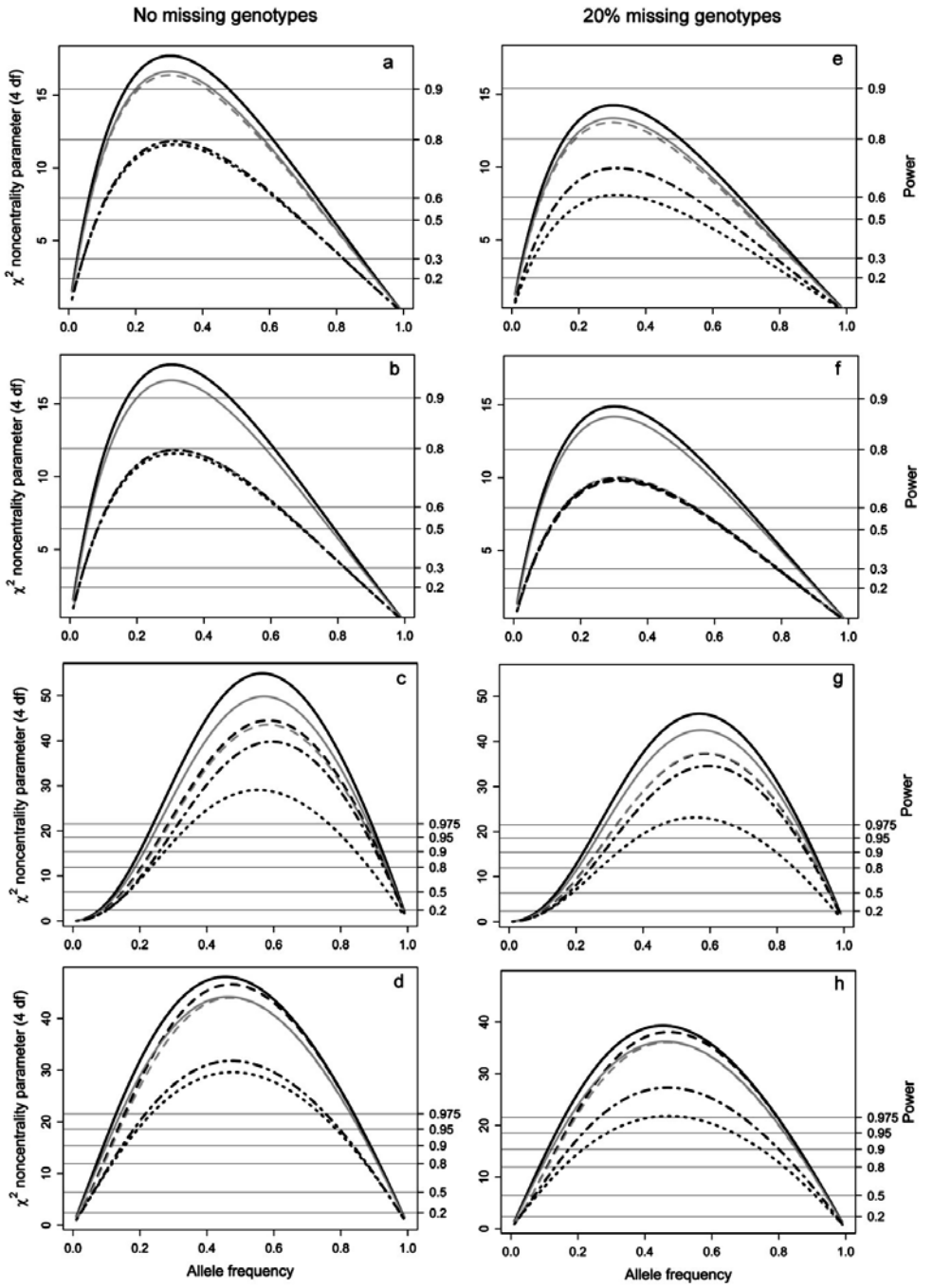
No missing genotypes
20% missing genotypes
a
b
c
d
e
f
g
h
χ² noncentrality parameter (4 df)
Power
Allele frequency
0.0
0.2
0.4
0.6
0.8
1.0
0.975
0.95
0.9
0.8
0.6
0.5
0.3
0.2

Figure 7.1 *Chi-squared noncentrality parameters and power as a function of allele prevalence for various designs and risk scenarios. Vertical axes: left, noncentrality parameter for a four-df likelihood ratio test of the null hypothesis R1=R2=S1=S2=1 based on 150 case families and 150 control families; right, power of corresponding test when* α*=0.05. Left column (panels a-d): no missing genotypes; right column (panels e-h): 20% missing genotypes. First row (panels a, e): R1=2, R2=3, S1=1, S2=1; second row (panels b, f): R1=1, R2=1, S1=2, S2=3; third row (panels c, g): R1=1, R2=2, S1=1, S2=3; fourth row (panels d, h): R1=1, R2=3, S1=2, S2=2. Designs: hybrid with control-mother dyads under mating symmetry (———) or mating asymmetry (– –); hybrid with control parents under mating symmetry (———) or mating asymmetry (– –); case-parent triads (·····); case-mother/control-mother (– · –). Curves that nearly or completely coincide: (·····, – · –) in panel a; (———, – –) in panels a, e; (– –, – –, ·····, – · –) in panels b, f; (– –, – –) in panels c, g; (———, – –) in panels a, d, e, h.*

With 150 case families and 150 control families, this population structure posed a serious challenge to the proposed hybrid design under model (1) by producing biased relative risk estimates and an inflated type I error rate for testing genetic effects (actual α level = 0.43 at nominal α = 0.05). Testing for the presence of bias due to population structure under mating symmetry at α = 0.10, the power was 0.42 and 0.54 for the four- and the one-df test, respectively, corresponding to noncentrality parameters of 3.71 and 3.03. Doubling the sample size doubles the noncentrality parameters, yielding power 0.69 and 0.79, respectively. All these values were near those achieved with the hybrid design using control parents where the analogous tests yielded power of 0.41, 0.53, 0.68 and 0.79, respectively[3].

We examined our ability to detect mating asymmetry for the proposed hybrid design using a sample size of 150 case families and 150 control families from a population with no population structure and null relative risks. For both the one-and the three-df tests, we verified that noncentrality parameters were zero under mating symmetry, as expected. We considered two specific scenarios where mating was not symmetric: one where only the three-df test would have power; a second where both the one-and the three-df tests would (Table 7.2). The patterns of asymmetry that we introduced for purposes of illustration were arbitrary and possibly unrealistic; we had no population-based data on mating asymmetry to guide us. For the mating-asymmetry scenario of Table 7.2, column 3, the asymmetry-induced bias for detecting relative risks using a model incorrectly assuming symmetry was small (actual α level = 0.055 at nominal α = 0.05). The one-df test had no power, as constructed; but the power of the three-df test was 0.90. For the asymmetry scenario of Table 7.2, column 4, the asymmetry-induced bias for detecting relative risks using a model incorrectly assuming symmetry was large (actual α level = 0.57 at nominal α = 0.05). The three-df test for asymmetry had power of 0.86, whereas the one-df test had power of 0.41.

Table 7.2 *Arbitrary mating-type configurations used to examine the power to detect mating asymmetry.*

Parameter value: proportion of families in each mating type			
Mating-type Parameter	**Under mating symmetry: HWE with p=0.5[a]**	**Under arbitrary mating asymmetry**	
		$\frac{1}{2}\mu_{01}+\frac{1}{2}\mu_{12}+\mu_{02} = \frac{1}{2}\mu_{10}+\frac{1}{2}\mu_{21}+\mu_{20}$	$\frac{1}{2}\mu_{01}+\frac{1}{2}\mu_{12}+\mu_{02} \neq \frac{1}{2}\mu_{10}+\frac{1}{2}\mu_{21}+\mu_{20}$
μ_{00}	0.0625	0.0625	0.0625
μ_{01}	0.1250	0.0950	0.0950
μ_{10}	0.1250	0.1550	0.1550
μ_{02}	0.0625	0.0875	0.0375
μ_{20}	0.0625	0.0375	0.0875
μ_{11}	0.2500	0.2500	0.2500
μ_{12}	0.1250	0.1050	0.1050
μ_{21}	0.1250	0.1450	0.1450
μ_{22}	0.0625	0.0625	0.0625

[a] Hardy-Weinberg equilibrium with allele frequency p=0.5.

Discussion

The popularity of family-based designs is due, in part, to their robustness to the insidious bias that genetic population structure can induce in population-based case-control comparisons. Population-based case-control designs, however, are superior for studying exposure effects on risk. The original motivation for combining family-based and population-based components was to use information from both sources to increase power for studying offspring genetic effects[4]. Family-based designs achieve robustness by conditioning on parental genotypes but thereby sacrifice information parental genotypes can offer about genetic risks. Any allele that increases risk in the offspring should be over-represented in the parents of cases compared to the parents of controls. Hybrid designs use their population-based component to provide additional information that can improve inference on genetic risks. The possibility that genetic population structure might bias a combined analysis could be addressed at best obliquely when the population-based component included only control subjects[5]. Hybrid designs that use control parents or control-mother dyads to enhance power offer a more straightforward ability to check for bias due to population structure while preserving the option of dropping back to the case-parents component when population structure is detected.

The primary motivation for considering a hybrid design that uses control-mother dyads instead of a hybrid design that uses parents of controls is a practical one: recruiting mothers of controls is easier than recruiting both mothers and fathers. Thus, larger

sample sizes should be easier to achieve and possible bias from selective refusals of fathers (or of mothers with concerns about paternity) can be reduced. On the other hand, control-mother pairs provide less direct information about mating-type parameters in the underlying population than do control-parent pairs. This feature is reflected in our results on power. The hybrid with control parents provided slightly more power for testing offspring and maternal genetic effects than did the hybrid with control-mother dyads.

In practice, the best design might be a hybrid of the two hybrid designs considered here. One could plan to recruit the parents of controls but, when the father is unavailable, try to recruit the control-mother dyad instead. Again, missing-data methods like the EM algorithm can be used for inference. Presumably the power for such a design would correspond to noncentrality parameters that fall between the curves of the two hybrid designs (Fig. 7.1). This hybrid-of-hybrids design would be tantamount to a design where case-parent triads are augmented by control-parent triads but missing data are dealt with in the analysis. The only advantage of genotyping control offspring when both parents are available, however, is the capacity to test the basic assumption of Mendelian transmission, an assumption that is not often questioned. For studying both offspring and maternal genetic effects on disease risk, hybrid designs that use control parents or control-mother dyads offer additional benefits over the case-parents design. The case-parents design is capable of studying the maternal genetic effects under the assumption of mating symmetry in the source population but offers no way to check that assumption. Hybrid designs provide information not only for checking that assumption but also for studying maternal genetic effects even when mating symmetry is not satisfied. These features were delineated previously for the hybrid that uses control parents[3], and we have shown here that they are maintained when control-mother pairs are substituted for control parents. Two different phenomena can masquerade as maternal genetic effects if not properly accounted for: mating asymmetry in the underlying population; and imprinting, where the effect of a transmitted allele depends on the parent of origin. After detecting a maternal effect, the investigator should ideally exclude the possibility that it was wholly or partly attributable to one of these other sources. With a case-parents design, mating symmetry in the underlying population must remain an assumption. The case-mother/control-mother design implicitly tolerates mating asymmetry. With either hybrid design, mating symmetry can be checked; and, if rejected, maternal effects can be studied under a model that accommodates mating asymmetry. Imprinting can be accommodated by the case-parents design using a log-linear model that incorporates parent-of-origin effects[13]. Although we have not considered it here, this same modeling tactic could be adapted for the hybrid designs or for the case-mother/control-mother design as well.

An appealing feature of hybrid designs that involve either parents of controls or

control-mother pairs is the ability to check key assumptions. The corresponding weakness is that these tests, particularly the key test for bias from population structure, appear to have limited power. For example, with 150 case and control families, we had little power to detect bias from population structure under a scenario which, if it went undetected, would induce substantial bias in inference. Power seemed somewhat better for detecting mating asymmetry in our examples; but we could certainly produce other examples where power was more limited. Larger sample sizes would help, of course. With a hybrid design, a simple practical precaution is to compare risk estimates under models that do and do not make a particular assumption. With concern about population structure, comparing relative risk estimates from the hybrid design with those from its case-parent triad component alone might provide reassurance. Similarly, with concern about mating symmetry, one could compare estimates from the mating-symmetry and the mating-asymmetry models.

The hybrid designs that we have considered will be useful for studies of complex phenotypes such as a birth defect where environmental factors will likely also play a role. The log-linear model for the 30-cell multinomial for case triads and control triads together can in principle be extended to incorporate environmental factors assessed categorically. Consequently, environmental risks and gene-by-environment interactions can be evaluated under either hybrid design. Postulating appropriate models for checking bias from population structure might be challenging, however, because exposure prevalence could change across genetically distinct subpopulations.

We have also demonstrated the potential increase in power that the hybrid designs can achieve when bias from population stratification is absent. The relative gain in efficiency (as measured by the ratio of the curves) depends on the underlying scenario. We based these power comparisons on equal numbers of case families in each design reflecting that the number of case families available is often a limitation when designing a study. The hybrid designs considered here require genotyping five individuals per case compared to only three or four, respectively, for the case-parents or the case-mother/control-mother designs. To compare designs when each of them genotyped 750 individuals (the number measured in the hybrid designs in Fig. 7.1), one would modify Figure 7.1 by multiplying the noncentrality curves for the case-parents design by 5/3 and those for case-mother/control-mother design by 5/4 (curves for hybrid designs remain the same). On that per-genotype basis, hybrid designs were not the most efficient in every scenario that we considered; however, they were more efficient than case-mother/ control-mother designs in general and more efficient than the case-parents design in realistic scenarios where genotypes were missing at random. In our view, any disadvantage in power per genotype of hybrid designs compared to the case-parents design is more than offset by the greater flexibility that hybrid designs offer for checking assumptions and evaluating effects of exposures. The cost of genotyping should not be the investigator's sole concern.

In summary, we have studied a hybrid design that combines data from case-parent triads and control-mother dyads. It provides the same flexibility as, though slightly less power than, a hybrid design that was introduced previously, one that uses the parents of controls instead of control-mother dyads. When needed but testable assumptions are met, either of these hybrid designs has a better power for studying both offspring and maternal genetic effects on disease risk than would a family-based case-parents design or a population-based case-mother/control-mother design.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (NIH Z01-ES040007) and by the Ter Meulen Fund. We thank Drs. Abee Boyles and Gregg Dinse for their helpful comments.

References

1. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet. 1998;62:969–978.
2. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of ''case-parent triads.'' Am J Epidemiol. 1998;148:893–901.
3. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet. 2005;77:627–636.
4. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet. 2004;12:964–970.
5. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet. 2005;76:592–608.
6. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet. 1993;53:1114–1126.
7. Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet. 1999;64:1186–1193.
8. Dempster AP, Laird NM, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977;39:1–38.
9. Sinsheimer JS, Palmer CGS, Woodward JA. Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. Genet Epidemiol. 2003;24:1–13.

10. Van den Oord EJ, Vermunt JK. Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads, by use of the computer program LEM. Am J Hum Genet. 2000;66:335–338.
11. Sun F, Flanders WD, Yang Q, Khoury MJ. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. Am J Epidemiol. 1999;150:97–104.
12. Agresti A. Categorical Data Analysis. 1990. New York: Wiley.
13. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet. 1999;65:229–235.

SHHM Vermeulen
M Den Heijer
P Sham
J Knight

Published in: Annals of Human Genetics 2007;71:1–12

CHAPTER 8

Application of multi-locus analytical methods to identify interacting loci in case-control studies

Abstract

To identify interacting loci in genetic epidemiological studies the application of multi-locus methods of analysis is warranted. Several more advanced classification methods have been developed in the past years, including multiple logistic regression, sum statistics, logic regression, and the multifactor dimensionality reduction method. The objective of our study was to apply these four multi-locus methods to simulated case-control datasets that included a variety of underlying statistical two-locus interaction models, in order to compare the methods and evaluate their strengths and weaknesses. The results showed that the ability to identify the interacting loci was generally good for the sum statistic method, the logic regression and MDR. The performance of the logistic regression was more dependent on the underlying model and multiple comparison adjustment procedure. However, identification of the interacting loci in a model with two two-locus interactions of common disease alleles with relatively small effects was impaired in all methods. Several practical and methodological issues that can be considered in the application of these methods, and that may warrant further research, are identified and discussed.

Introduction

Complex diseases or traits are a result of the interplay between and among genes and environmental factors. One way in which genetic factors can act together in relation to a certain trait is epistatically. Generally speaking, epistasis refers to the phenomenon that the relation between a gene and a trait outcome is dependent on another gene. More specifically speaking, from a statistical point of view, epistasis or gene-gene interaction has been defined as the presence of non-additivity in a mathematical model that describes the relation between genetic variants and a trait in a population[1,2].

The ability of multi-locus methods of analysis to detect and incorporate statistical gene-gene interaction in genetic association studies is likely to favour the chance of retrieving causal genes and a better prediction of trait outcome based on genotype data[1,3]. Researchers that want to identify interacting loci in their genetic association studies through application of multi-locus techniques can nowadays choose from a variety of methods and software packages that are readily available, in addition to traditional multiple regression analysis. The goal of our study was to apply four available multi-locus classification techniques, namely multiple logistic regression, the set association method as implemented in the Sumstat software, the multifactor dimensionality reduction (MDR), and logic regression, on simulated data to identify strengths and weaknesses in the application of these methods. More specifically, we have focussed on their ability to identify causal loci that may be statistically interacting according to some specific underlying model. This will give researchers more insight into the performance of the methods and allow more informed choices about when and how to apply them.

Many multi-locus methods that can be applied to in-vestigate gene-gene interactions have been introduced in the past decade, as reviewed by Hoh & Ott[4] and Thornton-Wells et al.[5]. A number of these methods are particularly suitable for genetic case-control association studies, and for some of those freely available software has been introduced. The MDR software was introduced in 2003 and was especially developed to analyse high-order joint effects of loci (and environment) in genetic association case-control and sib-pair study designs[6]. The general idea behind the non-parametric and genetic model-free MDR approach is that it reduces the dimensionality of the multi-locus data by pooling those combinations of genotypes that can be defined as high-risk and those that can be defined as low-risk, based on the case-control ratio for the specific multi-locus genotype. The reduction of the dimensionality of the data overcomes the problem of a low number of observations in high-order data combinations. An exhaustive search over all possible high-order genotype combinations for a varying number of loci is performed, and the best combination of single nucleotide polymorphisms (SNPs) for a certain model size is chosen based on classification error. Using cross-validation the ability of the new one-dimensional multi-locus variable to

predict case-control status, and in addition its cross-validation consistency, is determined. The locus or combination of loci that has the best predictive value and highest cross-validation consistency is considered to be the outcome of interest. An empirical p-value for the testing accuracy and cross-validation consistency of the final selected model can be determined using a permutation testing procedure. The MDR method has successfully identified interacting loci in several real data applications[7-12], and its performance has been evaluated on simulated data with a focus on a number of epistasis models that show no or small single locus main effects[11,13].

Logic regression is a generalized regression methodology that was introduced in 2001 and was also mainly developed to study high-order interactions in genetic studies[14,15]. It has been implemented in freely available software as an R and S-Plus package. In short, the goal of logic regression is to find Boolean combinations of binary predictors, for example SNPs, that are associated with the outcome of interest and that are incorporated into a regression model. The combinations of the binary predictors are efficiently organized in a tree form. The performance is evaluated by comparing the fitted values and response using a scoring function that is dependent on the type of regression. So instead of modelling the interaction between several SNPs using a high number of model parameters in a regression model, a combination can be captured through one tree parameter using Boolean operators; SNPs that are included in the same regression tree instead of in separate trees may be act-ing in a non-additive way on the scale of interest. The building and selection of the logic trees can be done using a stochastic simulated annealing algorithm, and the size and number of trees can vary. Model selection can be guided by comparing the predictive values of the models, based on cross-validation or permutation testing procedures. The latest software version also includes a Monte Carlo regression option that can identify several best tree models that are related to trait outcome and that may fit the data equally well[16]. The method has been applied to a post PTCA restenosis dataset and has also been successfully used in the analysis of simulated genetic data[14,15].

The set-association method, as implemented in the Sumstat software, introduced in 2001 by Hoh et al.[17], is a non-parametric method that uses sum statistics to evaluate the joint effect of loci related to trait outcome. The locus or set of loci associated with the trait of interest is identified by creating and testing sum statistics that capture the combined information from multiple SNPs. Based on single locus test statistics, for instance Chi-square values from 2 by 3 tables, Chi-square values for deviation from Hardy Weinberg equilibrium, or Chi-squares for pair-wise interactions, SNPs are selected and added to the sum statistic that is calculated for an increasing number of loci. The SNPs are chosen based on the value of the test statistic and are added sequentially to the sum statistic in order of decreasing value. The significance of each sum is evaluated using a permutation procedure. Then, the smallest empirical p-value for the sum statistic is in its turn evaluated for global significance via permutation tests, thereby correcting for the testing of multiple sums. So the goal is to find the subset of

loci that is most significantly associated with the trait of interest. This method has been evaluated for its power by use of simulated data without a focus on interaction[18,19], and has been applied in several case-control genetic association studies[17,20,21].

As well as several advantages, such as transparency, familiarity, and the estimation of interpretable effect parameters, the traditional multiple logistic regression technique has known disadvantages in identifying in-teracting loci in case-control association studies. Sparse data can easily become a problem in studies with relatively small sample sizes when including (high-order) interaction parameters. Furthermore, the number of tests can be very high when the number of SNPs included in the study, and the order of interactions to evaluate, becomes large. This can all lead to unreliable parameter estimates, inflated type I errors, and low power to detect the associated loci. Methods to deal with multiple comparisons like the Bonferroni correction, the False Discovery Rate (FDR), and randomization procedures, have been introduced as solutions to mitigate elevated type I error rates. Another difference from the previous discussed techniques is that the logistic regression is a parametric, model-based method that requires enumeration of the (interaction) model(s) to analyse; the available choice of model definitions that can be used in logistic regression to investigate gene-gene interactions is large. Logistic regression has successfully identified interacting loci in the past in studies that included a relatively small number of loci, and it was recently shown that it is also feasible to successfully apply the technique to identify interacting loci in genome-wide association studies[22].

The general interpretation of statistical interaction as deviance from additivity makes its presence dependent on the scale that is used (i.e. logit, or penetrance). The multiplicative model on a penetrance scale is often considered to represent the standard statistical interaction model; this model approximates an additive model on the log odds scale, implicit in the standard logistic regression model commonly used in the analysis of case-control studies. However, other statistical models besides the multiplicative also meet the criterion of non-additivity. In complex diseases the statistical interaction present in the collected population data is not known beforehand. It will then be important to know if the applied multi-locus methods can detect the interacting loci for different underlying interaction models. Therefore, we evaluated the performance of the above-mentioned multi-locus methods for a selection of two-locus statistical interaction models.

Materials and Methods

Statistical Interaction Models

We used the penetrance scale to define two-locus interaction models. The penetrance is the probability of being affected given the genotype at, in our case, both loci, and

it can range from 0 (no penetrance) to 1 (full penetrance). We simulated 5 two-locus interaction models with reduced penetrance based on the classification schemes described by Li & Reich[23] and those models described in earlier papers on epistasis (Figure 8.1). They include a 'multiplicative model' (Mod1) and a 'heterogeneity based model' (Mod2). The multiplicative is characterized by multi-locus penetrances that result from the product of the single locus contributions. The heterogeneity model is often considered a non-interactive model, in which both loci increase disease risk independent of the genotype at the other locus but it can also be viewed as a non-additive and therefore interactive model, as stated by Cordell[2].

Furthermore, we simulated a 'conditional dominant or recessive model' (Mod3), where one locus portrays dominant or recessive behaviour dependent on the genotype at the other locus. We also included an 'exclusive OR' (Mod4) and 'missing lethal genotype' (Mod5) based model. The first reflects the situation in which the risk of disease is elevated if a subject is heterozygous for one locus but not both. In the second model the disease risk is elevated when exactly two disease alleles are present, the risk being higher when both are at the same locus. These last two models are special because they show no main effects of the single loci in case of a disease allele frequency of 0.5.

We also simulated a model in which two two-locus interactions, both based on the 'missing lethal genotype' model (Mod5), increased disease risk according to a heterogeneity model (Mod6); the interaction was considered present if either one or the other or both sets of criteria for two-locus interactions were satisfied. The allele frequencies of both markers in both interactions were set to be equal. Note that this model also did not display any marginal single locus effects for the 0.5 frequency. Finally, a null model (ModNull) was generated in which the disease status of 200 out of the 400 randomly selected subjects was set to case.

Data Simulations

The data simulations were performed in Stata Version 8. We simulated the models by defining the penetrance and frequency for all two-locus genotype combinations for the variety of underlying disease models (Figure 8.1). We assumed that the two bi-allelic polymorphisms were both in Hardy Weinberg Equilibrium and in linkage equilibrium. Four different allele frequency models were generated, where both loci had an expected frequency of 0.5 (f0.5), 0.3 (f0.3), 0.1 (f0.1) or 0.05 (f0.05). In addition to the two causal SNPs eight non-causal SNPs were generated. Three of these had a frequency of 0.5, three of 0.3 and two had an allele frequency of 0.1.

To generate one simulated sample, 100 cases based on a high-risk genotype combination and 300 controls were randomly selected out of the complete simulated dataset with 150000 observations. To obtain a more realistic representation of a complex disease model, we introduced 50% phenocopies into every model by randomly

changing the disease status of 100 controls to case status thus creating datasets containing 200 cases and 200 controls. By doing this, we forced the disease prevalence of the population from which the final sample was selected to be twice that of the original simulations, since the 100 original true genetic cases now represented 50% of the total number of cases. As a result, the relative effect of the low frequency causal alleles was higher than that of the more frequent causal alleles. The final expected penetrances, the marginal penetrances and corresponding population prevalence of the disease are depicted in Table 8.1. For each combination of disease and allele frequency 100 replicates were simulated, which resulted in a total of 2800 datasets.

Mod 1

	BB	Bb	bb
AA	0	0	0
Aa	0	0.1	0.1
aa	0	0.1	0.1

Mod 2

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0	0.1
aa	0.1	0.1	0.1

Mod 3

	BB	Bb	bb
AA	0	0	0
Aa	0	0	0.1
aa	0	0.1	0.1

Mod 4

	BB	Bb	bb
AA	0	0.1	0
Aa	0.1	0	0.1
aa	0	0.1	0

Mod 5

	BB	Bb	bb
AA	0	0	0.1
Aa	0	0.05	0
aa	0.1	0	0

Figure 8.1 *Two-locus penetrances for the models used for data simulation.*

Statistics

Multiple Logistic Regression Analysis

The parametric multiple logistic regression technique requires specification of the statistical model and parameters that we want to evaluate for association with disease. We chose to fit a full logistic regression model for each pair of loci. The loci were coded using two dichotomous dummy variables: one for the heterozygous group and one to indicate the group homozygous for the high-risk allele. The fully fitted logistic regression model therefore contained an intercept, four parameters for the main effects, and four parameters for the interaction terms between the two loci. We applied both the Bonferroni (fixed overall type I error to 0.05) and the FDR method (Simes procedure, FDR 0.05) to deal with the multiple comparisons. Those single loci or interaction

parameters that passed the significance criterion were considered our outcomes of interest. The logistic regression analysis was performed in Stata Version 8; we used the 'multproc' add-on package written by R. Newson for the Bonferroni and FDR procedure.

Table 8.1 *Population genotypic penetrances for the simulated case-control data after introduction of the phenocopies.*

											marginal penetrance[3]			
Model	MAF[1]	f_{00}[2]	f_{01}	f_{02}	f_{10}	f_{11}	f_{12}	f_{20}	f_{21}	f_{22}	$f_{0.}$	$f_{1.}$	$f_{2.}$	K[4]
1	0.5	0.06	0.06	0.06	0.06	0.15	0.15	0.06	0.15	0.15	0.06	0.13	0.13	0.11
	0.3	0.03	0.03	0.03	0.03	0.12	0.12	0.03	0.12	0.12	0.03	0.08	0.08	0.05
	0.1	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.10	0.10	0.00	0.02	0.02	0.01
	0.05	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.10	0.10	0.00	0.01	0.01	0.00
2	0.5	0.05	0.05	0.14	0.05	0.05	0.14	0.14	0.14	0.14	0.07	0.07	0.14	0.09
	0.3	0.02	0.02	0.12	0.02	0.02	0.12	0.12	0.12	0.12	0.03	0.03	0.12	0.03
	0.1	0.00	0.00	0.10	0.00	0.00	0.10	0.10	0.10	0.10	0.00	0.00	0.10	0.00
	0.05	0.00	0.00	0.10	0.00	0.00	0.10	0.10	0.10	0.10	0.00	0.00	0.10	0.00
3	0.5	0.03	0.03	0.03	0.03	0.03	0.13	0.03	0.13	0.13	0.03	0.06	0.10	0.06
	0.3	0.01	0.01	0.01	0.01	0.01	0.11	0.01	0.11	0.11	0.01	0.02	0.06	0.02
	0.1	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.10	0.10	0.00	0.00	0.02	0.00
	0.05	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.10	0.10	0.00	0.00	0.01	0.00
4	0.5	0.05	0.15	0.05	0.15	0.05	0.15	0.05	0.15	0.05	0.10	0.10	0.10	0.10
	0.3	0.05	0.15	0.05	0.15	0.05	0.15	0.05	0.15	0.05	0.09	0.11	0.09	0.10
	0.1	0.03	0.13	0.03	0.13	0.03	0.13	0.03	0.13	0.03	0.05	0.11	0.05	0.06
	0.05	0.02	0.12	0.02	0.12	0.02	0.12	0.02	0.12	0.02	0.03	0.11	0.03	0.03
5	0.5	0.03	0.03	0.12	0.03	0.07	0.03	0.12	0.03	0.03	0.05	0.05	0.05	0.05
	0.3	0.02	0.02	0.12	0.02	0.07	0.02	0.12	0.02	0.02	0.03	0.04	0.07	0.04
	0.1	0.00	0.00	0.10	0.00	0.05	0.00	0.10	0.00	0.00	0.00	0.01	0.08	0.01
	0.05	0.00	0.00	0.10	0.00	0.05	0.00	0.10	0.00	0.00	0.00	0.01	0.09	0.00
6[5]	0.5	0.06	0.06	0.12	0.06	0.09	0.06	0.12	0.06	0.06	0.08	0.08	0.08	0.08
	0.3	0.05	0.05	0.12	0.05	0.08	0.05	0.12	0.05	0.05	0.05	0.06	0.08	0.06
	0.1	0.01	0.01	0.10	0.01	0.06	0.01	0.10	0.01	0.01	0.01	0.02	0.09	0.01
	0.05	0.00	0.00	0.10	0.00	0.05	0.00	0.10	0.00	0.00	0.00	0.01	0.09	0.00

1 MAF = minor allele frequency.

2 f_{ij} represents the penetrance for a two-locus genotype where the number of disease alleles at the first locus is i and at the second locus is j.

3 marginal penetrance defined by $f_{i.}=\sum p_j f_{ij}$ where p_j is the frequency of genotype j.

4 K = disease prevalence.

5 only marginal penetrances are displayed; here f_{ij} represents the marginal penetrance of the ij.. multi-locus genotype.

Set-association method

The set-association method as implemented in the Sumstat software available at http://linkage.rockefeller.edu/register was applied. The Chi-square statistic for genotypic association with case-control status (2 by 3 table) was chosen as the test statistic. Furthermore, we selected the interaction option that computes the Chi-squares for all pair-wise interactions, leading to a total number of 55 genotypic input variables for every replicate. We calculated sum statistics that contained at most 6 variables (single loci and/or interactions) and used 20000 permutations to generate empirical p-values for the test statistics. For every replicate the subset of single loci and pair-wise interaction terms with the lowest permuted p-value and largest number of terms was the multi-locus combination output of interest.

Logic Regression

We performed the analysis using the Logic Regression module that is available as an R package (version 1.3.1; http://bear.fhcrc.org/~ingor/logic/), and selected the logistic regression option. We used 100,000 iterations in the simulated annealing algorithm, and annealing temperatures were set so the number of acceptions/rejections during the iteration process was optimal. The logic regression can only deal with dichotomous input variables, and therefore the SNPs were recoded into two dichotomous dummy variables as was done for the logistic regression analysis. We used the 'select the best model' option and limited the model size to a maximum of 4 and 6 leaves for the two-locus models, and 8 leaves for the 4-locus model. The outcome of interest was the loci present in the final best model, as selected under our model selection settings.

MDR

We performed the analysis using MDR version 0.6.1, available at http://www.epistasis.org/open-source-mdr-project.html. We evaluated model sizes of 2 and 3 loci and 4 loci for the 4-locus genetic heterogeneity models, and counted the number of replicates that contained the causal loci in the final best model that was identified using 10-fold cross-validation. The default values for the parameters that needed to be set were used; the threshold case/control ratio for the definition high-risk genotypes was set to at least 1:1, tie cells were set to affected, and an exhaustive search method configuration was selected.

Results

Figure 8.2 shows the number of replicates that contained the two causal loci, either both loci separately or via an interaction, for all the scenarios and methods. For the logistic regression and Sumstat results we could indicate the number of replicates in

which the loci were identified via an interaction term, and these therefore would have been positive replicates if we would had selected on interaction terms only (black shading). For the genetic heterogeneity model, with two two-locus interactions acting to increase risk of disease, we counted the number of replicates in which the four loci were retrieved, either by identifying the simulated interaction terms or the single locus effects (Figure 8.3). Because of the low penetrance and low frequency of the risk increasing genotypes for Mod2f0.05, Mod3f0.1 and Mod3f0.05, the final simulated case-control datasets contained too few cases (<100) and were therefore dropped.

For the traditional logistic regression technique, without correction for multiple testing, the interacting loci were identified in almost all replicates, even when the causal allele frequency dropped to 0.05. The detection rate was diminished for the four-locus models with a frequency of 0.3 and 0.5. After application of the FDR procedure and Bonferroni method, the number of replicates in which the causal loci were detected was somewhat less in most models, but seriously impaired in the models with an allele frequency of 0.5 and the four-locus models, the Bonferroni correction being more conservative than the FDR method. The loci were found specifically through the interaction terms in Mod4f0.5, Mod5f0.5 and Mod6f0.5, where there were no expected marginal effects. Only relying on the interaction parameters generally resulted in lower detection of causal loci, and failed especially in Mod2. It was also worse for the 0.05 MAF models where the interaction terms were dropped due to sparseness of data. Furthermore, we saw that the number of replicates with significant interaction parameters was lower for the FDR corrected results and even lower for the Bonferroni results, in most scenarios.

The number of false positives for the null model in the uncorrected analysis was highly inflated. There was a high false positive detection rate for the null model with a frequency of 0.1 that could not be mitigated by application of the multiple comparison procedures. Looking at the results and analysis more carefully showed that the sparse number of observations in the genotype cells for this allele frequency led to unreliable parameter estimates, and a high rate of extreme low p-values that passed the significance criteria, whereas the parameters of the causal loci were often dropped and not tested in the 0.05 frequency model, due to too few observations in the heterozygote, but especially homozygous, mutant genotype groups. Likewise, the interaction terms were dropped, explaining their absence in the null model findings.

In general, Sumstat performed very well for the 0.3, 0.1 and 0.05 MAF scenarios and not well for Mod1f0.5, Mod2f0.5 and high frequency Mod6. Regarding the last model, in Figure 8.3 we only counted those replicates as positive if the correct single loci and/or both simulated interaction terms were identified. We saw, however, that in the lower frequency models the four causal loci were identified via interaction terms between one locus from the first two-locus interaction and one locus from the second. If we had counted those replicates as positive results too, the number of replicates

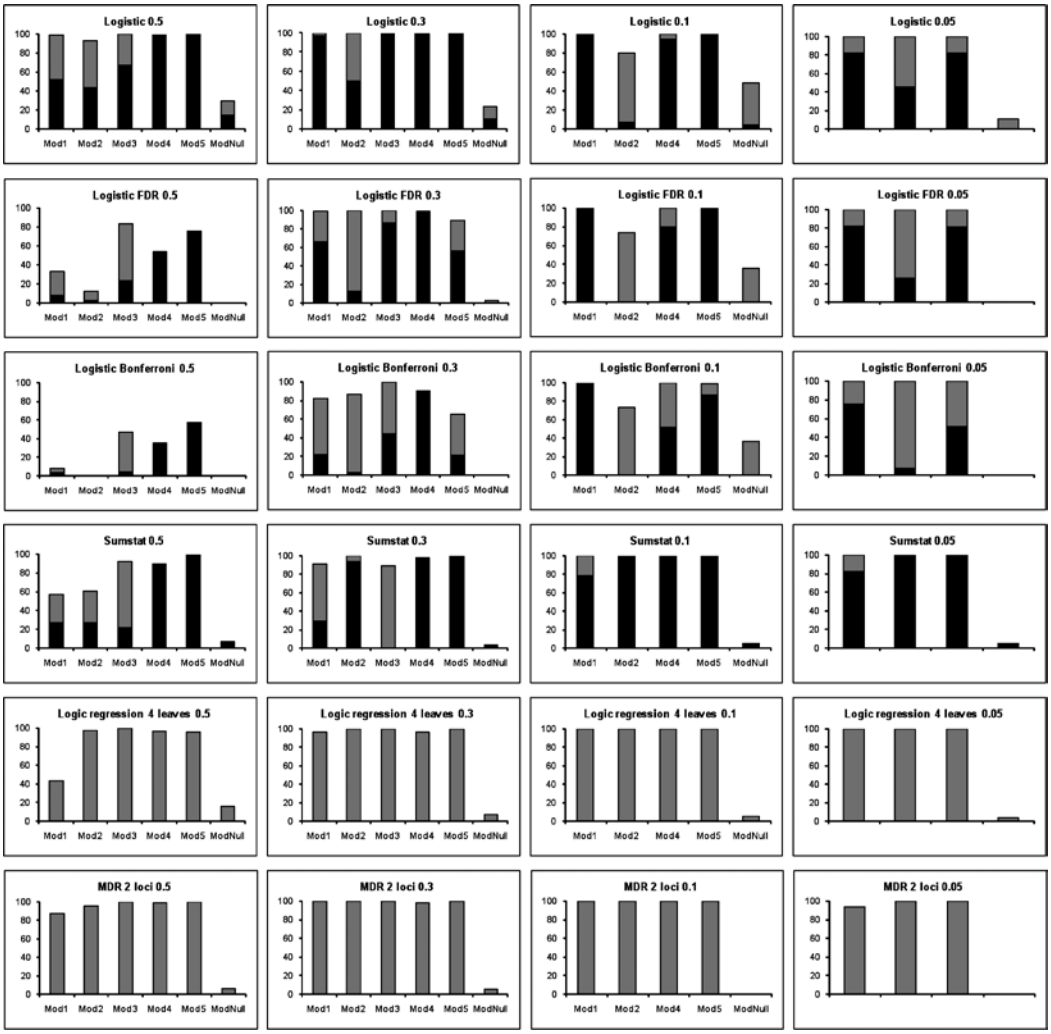

Figure 8.2 *Number of replicates that contained the two causal loci via both single locus parameters and/or through interaction parameters (black shading) for the 5 two-locus models and the null model. The first to the fourth column represent the models with minor allele frequencies of 0.5, 0.3, 0.1 and 0.05, respectively.*

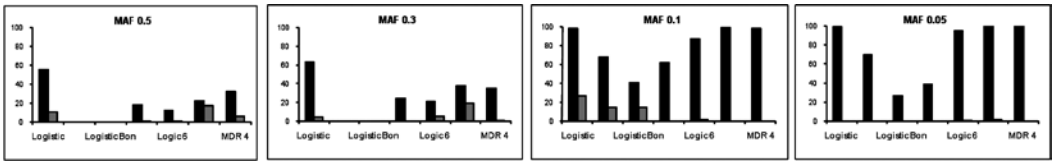

Figure 8.3 *Number of replicates per method that contained the four causal loci in Mod6 (black shading) and ModNull (grey shading) for every MAF.*

in Figure 8.3 for the 0.1 and 0.05 frequencies would have been 100 and the results for the 0.5 and 0.3 frequencies would have increased to 22 and 32, respectively. For all frequencies the Sumstat method retrieved the causal loci for Mod4 and Mod5 via the interaction term, especially for the 0.1 allele frequency, and Mod3 was mainly dependent on finding the separate single locus effects. The results for both single causal loci (not shown) were expected to be symmetrical, but for some models they differed substantially up to a maximum of 19 replicates for Mod1f0.5. We saw that for this genetic model all methods gave asymmetrical results for the two causal loci; the first locus was retrieved more often than the second. The overall low number of falsely detected loci, as observed in the null model, was low relative to the other methods, even without selection on the basis of the empirical sum statistic p-values or the final global p-value. In the ModNull we saw a higher selection of interaction terms over single terms.

We also checked how the results would change when only the replicates with a global p-value of less than 0.05 were considered. The number of false positives in the null models was reduced to zero. It had marginal consequences for the results of Mod1f0.5, Mod2f0.5, Mod4f0.5, where the number of replicates containing the true loci dropped at most 7, and had large consequences for Mod3f0.05 and Mod5f0.5/f0.3, where the number of replicates that contained the causal loci was lowered to 16, 49 and 28, respectively. For the four-locus models the numbers in Figure 8.3 changed for the f0.5 and f0.3 models to 4 and 14, respectively.

The results were different if the sum statistic with the smallest number of loci for the smallest p-values, instead of the largest number, was chosen (data not shown). We then saw that the method was much worse for Mod1, Mod2f0.3 and Mod3 in terms of elevated type I error, because the p-value often reached its minimum of zero after including one of the two causal loci. The method was also not able to identify the four loci in Mod6f0.1 and Mod6f0.05.

The logic regression analysis retrieved a high number of replicates containing the true causal loci, with the exception of the high frequency multiplicative model (Mod1f0.5) and Mod6, where it had difficulties finding all 4 loci. The amount of noise varied; Mod1 and Mod3 in general contained at least 3 terms related to the two causal loci, with the exception of Mod1f0.5 and Mod2 which contained just two terms referring to the causal loci and two referring to non-causal loci. The lower frequency Mod5 often contained 4 causal loci terms, and all Mod4 contained 2 to 4 causal terms out of the four leaves. In all scenarios one tree was constructed, often with the maximum number of leaves allowed. The gain from allowing six leaves instead of four was very limited, and led to a higher inclusion of non-causal loci (results not shown). The smaller the model, that is the lower the number of leaves allowed, the lower the number of false positive findings. We observed a tendency for a lower number of false positive findings for low frequency SNPs, reflected in the decreasing values for ModNull from f0.5 to f0.05.

The high frequency four locus Mod6 contained a high number of non-causal loci and a varying number of true loci. The performance for the low frequency scenarios was remarkably better. The difference between the analysis allowing 6 and 8 leaves was most pronounced for the 0.3 and 0.5 MAFs, where we saw a higher number of correct positive findings but also much higher false positive results for the analysis with 8 leaves.

The same trends in results were observed for the MDR as for the logic regression: it retrieved the loci in all situations, except for the high frequency Mod6, and the performance was slightly impaired in Mod1f0.5. Because the model could only contain two loci, no non-causal loci were included in those models that correctly identified the loci. There was no additional value from models containing three instead of two loci; the retrieved number of causal loci remained similarly high but the number of false positives was substantially higher in the high frequency models. Also here we observed a bias towards high frequency SNPs for the null model.

Discussion

We performed a simulation study to identify strengths and weaknesses of four multi-locus methods employed to identify interacting loci in several case-control scenarios. The number of ways in which these methods can be applied is numerous. Due to pragmatic and computational reasons we have not used the methods to their full capacity regarding model selection and model testing options. Therefore we can only draw conclusions about the strengths and weaknesses we encountered in the ways we applied them to the simulated data. Since we did not fix the type I error for all methods in an equivalent way it is not possible to make fair direct comparisons of power between the methods, and we have focussed on identifying strengths and weaknesses for each method separately.

In the application of the logistic regression the known problems in standard regression techniques of inflated findings of false positives, and diminished power caused by the presence of sparse data and multiple testing problems, were encountered, despite there being only 10 loci in these datasets. We can therefore underscore the importance of the use of a model testing procedure, such as permutation tests, even in studies with few genetic variants. Furthermore, extensions of the traditional regression models, that are especially developed to deal with sparse data, like penalized regression models could be applied. We chose to apply a full model containing parameters for the two loci and their interaction terms. We observed that the significance of the interaction parameters in a fully saturated two-locus logistic regression model in the identification of the causal loci was dependent on the underlying genetic models, the allele frequency, and the multiple comparison correction procedure. The use of interac-

tion parameters especially improved the identification of interacting loci in cases of statistical interaction models that showed no expected marginal effects. North et al.[24] evaluated how well logistic regression models that included different model parameters, including interaction parameters, fitted the data compared to a fully saturated model, and how well they corresponded to the true underlying penetrance based models for a variety of two-locus disease models. They showed that for some models the inclusion of interaction parameters is advantageous but there is no direct correspondence between the interactive effects in the logistic regression models and the underlying penetrance based models displaying some kind of epistasis effect[24]. The latter has been confirmed in our study.

An approach that we have not evaluated here but that can be applied to identify gene-gene interactions, is the case-only approach. It can be used to efficiently identify deviation from a multiplicative model for the relative risks by testing the association between loci in the cases. This design can be more powerful than the traditional case-control analysis. However, it is restricted to the identification of interactions only; the main effects of loci cannot be estimated and tested. Furthermore, this approach is sensitive to deviation of the underlying assumption of independence of the loci in the general population[25]; bias will be introduced if linkage disequilibrium between the loci of interest exists in the control population.

The Sumstat method had difficulty finding the causal loci in the high frequency four-locus models, but for the other scenarios the results were good. We have not performed the set-association method without construction of the two-locus interaction input variables, and it is therefore not possible to discuss the added value of using these in addition to the single locus parameters as a test statistic. We did see that the identification of the loci through the interaction parameter or single loci variables was dependent on the underlying model and the MAF, but the loci were consistently found via the interaction term in the 'exclusive OR' and 'missing lethal genotype' scenarios. The inability to deal with interacting loci that show no or weak main effects is an often mentioned disadvantage of the set association approach[26]. Heidema and colleagues state in their review that genetic interactions are only tested for the loci that are incorporated into the sum statistic. The fact that loci are incorporated into the sum statistic does, in our view, not deal with interactions between these loci since the separate test statistics are simply added, but we saw that this can be overcome by introducing the interaction test statistics prior to calculating the sum statistics. The current Sumstat version is not equipped to handle high-order interactions; only two-locus interaction parameters can be constructed. This was not a problem in our simulations, but could be a shortcoming in real usage. Selection of the results based on the permutation global p-values reduced the type I error perfectly. However, this was at the expense of a large loss of power to retrieve the causal loci for some models.

The results for the logic regression and MDR were similar. They performed well

for all models, with the exception of the high frequency heterogeneity models with two different two-locus interactions, and showed a slightly impaired performance for the high frequency models with relatively small effect sizes. One disadvantage of our procedure is that we did not apply model selection techniques other than limiting the number of genetic parameters in the analysis. This meant that we could not evaluate the underestimation of the type II error and overestimation of the type I error in our study that was caused by not selecting the optimal model size and not using global p-value cut-off points to select the outcome of interest.

One of the strengths of the MDR that is often highlighted, is its high power to identify high-order interactions for loci without a main effect[26]. Our results confirm this for two-locus interactions. We also saw that this characteristic was not only limited to the MDR method, as the logic regression and Sumstat were capable of identifying these loci too. We have however not compared their power for a fixed type I error.

The number of replicates containing the loci under the null model decreased when applying a simple model size limiting strategy, by maximizing the number of leaves and loci, respectively. Limiting the model size to the smallest possible size considering the number of true causal loci was advantageous compared to a larger model. However, restraining the model size to a number lower than the actual number of causal SNPs present in the data obviously impaired the performance of the methods. The publicly available MDR software is currently limited to a maximum of 15 loci. When a larger number of loci might be involved in the genetic aetiology this limit would be too small.

There was a tendency in both methods to show a lower number of positive replicates under the null model when the frequency of the SNPs of interest decreased. This bias towards high frequency SNPs might warrant follow-up study to investigate more carefully the causes and consequences of this preference.

To judge if the loci included in the multi-locus combination act additively or not, the user of the logic regression and MDR method can evaluate the number of trees and the way the loci are included in the trees, and the graphically displayed case-control ratios for all multi-locus genotype combinations, respectively. This can become a daunting task, especially for the MDR method, when the number of loci included increases, and also when the number of datasets is large as in our case. We therefore did not explore this option in the current study, but we did find that for the logic regression all loci were consistently combined in one tree, pointing towards non-additive association between the loci and the case-control status in the logistic regression.

Finding similar results for the MDR method and the logic regression was not surprising. The similarities in methodology between the MDR and an extension of a recursive partitioning technique related to the logic regression approach were discussed by Bastone et al.[27]; the MDR method can be viewed as a special form of recursive partitioning technique[27]. One major distinction between the MDR and logic regres-

sion is that in the first approach tree growth is restricted to a single split in one tree. Based on this fact one might expect the logic regression to perform better under genetic heterogeneity because of its ability to create multiple splits and more than one tree. Also Heidema et al.[26] state that recursive partitioning techniques, of which logic regression is one example, should be able to detect genetic heterogeneity. This was not confirmed in the results of this study, where the logic regression had difficulty identifying the causal loci for some four-locus models. The two-locus heterogeneity based model was tackled well by the logic regression as well as by the MDR method. Ritchie et al.[13] identified the impaired performance under genetic heterogeneity of the MDR method. In their simulations the power of the MDR for a model comparable to our four-locus model was worse, probably reflecting the selection on the statistical significance of the output they used. The authors propose the use of cluster analysis or recursive partitioning techniques, to identify clusters of individuals with similar genetic backgrounds prior to performing the MDR to mitigate the low power in genetic heterogeneity, and state that more research into these methods is needed.

We applied the methods on a large number of different statistical interaction models for a variety of allele frequency possibilities, but our simulations were not exhaustive. The high frequency models with two two-locus interactions turned out to be the most challenging. They were not only characterized by interacting risk factors and genetic heterogeneity, but also by small expected marginal effects. So even though, in the individual situations of a statistical heterogeneity model or an underlying penetrance model with small marginal effects, the causal loci were retrievable, in the combined situation power was diminished. Furthermore, it is obvious that we have not set out to explore the limits regarding model complexity, effect size, allele frequency, sample size, and their combinations for these methods. The authors of the methods have touched upon these aspects. Future research could explore the performance of the methods for more complex underlying models with several one-locus effects and high-order interactions, gene-environment interactions and covariates.

We have simulated data for a small number of SNPs in order to evaluate the performance of multi-locus methods regarding the identification of interacting loci. With the increasing application of large candidate gene arrays and genome-wide SNP arrays it will be of importance to evaluate the strengths and weaknesses of multi-locus methods for handling large amounts of genetic data where linkage disequilibrium is likely to be present. The specific relative strengths and weaknesses in handling a large amount of SNP data for a diversity of methods, including logistic regression, the set association approach and the MDR, have been recently discussed by Heidema et al.[26].

Ideally, one would limit the analysis of genetic association data to one or a few methods and apply them in a way that would enable capture of all the underlying signals from associated genes, whether acting singularly or interactively, and give insight into the underlying complex aetiology of the disease or trait of interest. Since every

method has its own strengths and weaknesses, and within every method a diversity of approaches can exist, a multi-analytic approach could help in distinguishing between true and false positive findings. It is however important to apply the methods in the most optimal way, and to understand the limits of each method in order to correctly interpret the results. The results of this study can give researchers insights into how to apply the discussed methods best in practice, judge where they perform similarly, and help in interpreting the results of the different methods.

Acknowledgements

This study was supported by the Netherlands Heart Foundation, Grant 2002B68. Sita Vermeulen has received a Frye Stipend and was supported by a travel grant from the Stichting Simonsfonds. Martin den Heijer is supported by a VENI grant from the Netherlands Organisation for Scientific Research (NWO). Jo Knight has an MRC Bioinformatics Training Fellowship, Grant G0501329. Pak Sham was supported by National Eye Institute Grant EY-12562, Hong Kong Research Grants Council CERG Grant HKU7669/06M and The University of Hong Kong Strategic Research Theme on Genomics, Proteomics and Bioinformatics. We would like to thank S. Newman for performing the Sumstat analysis.

References

1. Cordell HJ, Todd JA, Hill NJ, Lord CJ, Lyons PA, Peterson LB, Wicker LS, Clayton DG. Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. Genetics. 2001;158:357–367.
2. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002;11:2463–2468.
3. Culverhouse R, Suarez BK, Lin J, Reich T. A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet. 2002;70:461–471.
4. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet. 2003;4:701–709.
5. Thornton-Wells TA, Moore JH, Haines JL. Genetics, statistics and human disease: analytical retooling for complexity. Trends Genet. 2004;20:640–647.
6. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003;19:376–382.
7. Brassat D, Motsinger AA, Caillier SJ, Erlich HA, Walker K, Steiner LL, Cree BA, Barcellos LF, Pericak-Vance MA, Schmidt S, Gregory S, Hauser SL, Haines JL, Oksenberg JR, Ritchie MD. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. Genes Immun. 2006;7:310–315.

8. Cho YM, Ritchie MD, Moore JH, Park JY, Lee KU, Shin HD, Lee HK, Park KS. Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. Diabetologia. 2004;47:549–554.
9. Coffey CS, Hebert PR, Ritchie MD, Krumholz HM, Gaziano JM, Ridker PM, Brown NJ, Vaughan DE, Moore JH. An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC Bioinformatics. 2004;5:49.
10. Ma DQ, Whitehead PL, Menold MM, Martin ER, Ashley-Koch AE, Mei H, Ritchie MD, DeLong GR, Abramson RK, Wright HH, Cuccaro ML, Hussman JP, Gilbert JR, Pericak-Vance MA. Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. Am J Hum Genet. 2005;77:377–388.
11. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69:138–147.
12. Williams SM, Ritchie MD, Phillips JA 3rd, Dawson E, Prince M, Dzhura E, Willis A, Semenya A, Summar M, White BC, Addy JH, Kpodonu J, Wong LJ, Felder RA, Jose PA, Moore JH. Multilocus analysis of hypertension: a hierarchical approach. Hum Hered. 2004;57:28–38.
13. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003;24:150–157.
14. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet Epidemiol. 2001;21Suppl1:S626–S631.
15. Ruczinski I, Kooperberg C, LeBlanc L. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression, with applications. J Multiv Anal. 2004;90:178–195.
16. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol. 2005;28:157–170.
17. Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res. 2001;11:2115–2119.
18. Kim S, Zhang K, Sun F. Detecting susceptibility genes in case-control studies using set association. BMC Genet. 2003;4Suppl1:S9.
19. Wille A, Hoh J, Ott J. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol. 2003;25:350–359.
20. de Quervain DJ, Poirier R, Wollmer MA, Grimaldi LM, Tsolaki M, Streffer JR, Hock C, Nitsch RM, Mohajeri MH, Papassotiropoulos A. Glucocorticoid-related genetic susceptibility for Alzheimer's disease. Hum Mol Genet. 2004;13:47–52.
21. Maitland-van der Zee AH, Turner ST, Schwartz GL, Chapman AB, Klungel OH, Boerwinkle E. A multilocus approach to the antihypertensive pharmacogenetics of hydrochlorothiazide. Pharmacogenet Genomics. 2005;15:287–293.
22. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005;37:413–417.

23. Li W, Reich J. A complete enumeration and classification of two-locus disease models. Hum Hered. 2000;50:334–349.
24. North BV, Curtis D, Sham PC. Application of logistic regression to case-control association studies involving two causative loci. Hum Hered. 2005;59:79–87.
25. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. Am J Epid. 2001;154:687–693.
26. Heidema GA, Boer JM, Nagelkerke N, Mariman EC, Van Der A DL, Feskens EJ. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. BMC Genet. 2006;7:23.
27. Bastone L, Reilly M, Rader DJ, Foulkes AS. MDR and PRP: a comparison of methods for high-order genotype-phenotype associations. Hum Hered. 2004;58:82–92.

CHAPTER 9

General discussion

This chapter includes a discussion of the studies described in this thesis. The contribution of our studies in light of our objectives and recent findings in other studies are discussed in section 9.1. Next, in section 9.2, we highlight some of the strengths and weaknesses related to the study designs and utilized study populations, measurement of the main variables, and statistical analyses that have affected our ability to reach the objectives of this thesis. Section 9.3 discusses perspectives for future research.

9.1 Contribution of the studies in light of our objectives and recent findings in other studies

*Large number of small-scale and small number of large-scale genetic association studies for plasma tHcy published in recent years: nothing beats the MTHFR*677C>T

Before the start of this project, genetic epidemiological research into plasma tHcy had generally focused on a limited number of nonsynonymous variants in candidate genes in small scale association studies. The studies resulted in the conclusive identification of the *MTHFR*677C>T as determinant of plasma tHcy as well as folate concentrations and inconclusive findings for other potential genetic determinants including the 31 bp VNTR in *CBS*[1] (see also Table 1.1). A MEDLINE search for English reports published between January 2005 and August 2008 using the terms [(homocysteine) AND (gene)] yielded 707 studies; hand-searching of the available abstracts indicated that 179 papers contained information on DNA variants as well as plasma tHcy concentrations. Mostly, these concerned small (n<400) to moderate-size (n=400-1000) studies that included diseased and/or healthy individuals and in which the association between previously described nonsynonynous DNA variants for plasma tHcy as well as disease was evaluated. A number of these genetic analyses have been performed in large-scale (n>1000) epidemiological studies in which plasma tHcy and additional one-carbon metabolites have been measured. These include the 'Norwegian Colorectal Cancer Prevention (NORCCAP)' study[2,3], the 'Hordaland Homocysteine study'[4], the 'Cardiovascular Risk in Young Finns Study'[5], the 'Health in Men study'[6], the 'National Health and Nutrition Examination Survey DNA Bank (NHANES) III'[7], and the 'Women's Health Study'[8].

The *MTFHR*677C>T was by far the mostly studied DNA variant for which predominantly positive associations with increased plasma tHcy were reported, especially among the moderate-size[9-25] and large studies[2,3,5,7,8,26-34]. The second most studied DNA variant was *MTHFR*1298A>C with mostly negative (e.g. [7,22]) but also positive results (e.g. [3,25]) regarding association with plasma tHcy. However, in allmost all of these studies the correlation between the *MTHFR*1298A>C and *MTHFR*677C>T was not accounted for which may have resulted in false positive findings for *MTHFR*1298A>C. Also *MTR*2756A>G and *MTRR*66A>G were measured by multiple research groups, again

with contradicting association results, also among the moderate and large-scale studies[3,7,11,18,32,35]. Other DNA variants that are displayed in Table 1.1 were evaluated in at most eight studies. Devlin et al.[30] did not find an influence of the *FOLH1*1561C>T variant on plasma tHcy concentrations, in contrast to others[4,36]. Reports for *SLC19A1*80G>A and plasma tHcy were predominantly negative but associations with serum folate were reported[3,30,36-40]. Seven studies generally refuted a relation for *TCN2*776C>G and plasma tHcy[3,11,40-44] and two reports on the *TCN2*67A>G did not find association to plasma tHcy[3,40]. Two large-scale studies (including 6,793 and 10,601 subjects, respectively) that have measured the 68 bp insertion/deletion polymorphism in *CBS* showed contradicting results for association to plasma tHcy[3,7]. An association between the 31 bp VNTR in *CBS* to plasma tHcy concentrations, previously reported by our group in 2001[45], has been replicated in 1,400 subjects of the Framingham population[46]. No other studies on this variant have been published yet. Also, no association studies for plasma tHcy were found for the *CTH*1364G>T variant.

Publications concerning DNA variants other than those already presented in Table 1.1 were relatively sparse. The *COMT*324A>G has been studied by multiple groups in the last eight years, including our group, and showed contradicting results for plasma tHcy concentrations[47-51]. Three reports on the −786T>C variant in *eNOS* were found among the 179 selected studies; two studies showed a positive association between this variant and plasma tHcy[26,32] but this was not confirmed in a third study[52]. In 2005, Souto et al.[53] identified a plasma tHcy-affecting haplotype, including SNP rs694539, in *NNMT* via a linkage analysis approach. However, one follow-up study refuted an association for rs694539 in *NNMT* and plasma tHcy in their population[35] and a second association study showed inconclusive results[54]. *PON1*163T>A and *PON1*575A>G, two nonsynonymous SNPs in the paraoxonase 1 gene, were studied by Fredriksen et al.[3] and did not show association to plasma tHcy. Both *PON1* polymorphisms were also measured by Shin et al.[55]; in contrast, they did find an influence of *PON1*163T>A on plasma tHcy. Golledge and Norman[6] were the only ones to report on 34G>C and 1347C>T in the peroxisome proliferator-activated receptor gamma gene (*PPARG*) and showed convincing association to plasma tHcy for both variants in their population consisting of 3,875 elderly men. The association between a 19 bp deletion in intron 1 of *DHFR* that our group reported in 2007[56] was not replicated by a recent study by Stanislawska-Sachadyn and colleagues[57].

Most studies reported on one to at most five different genetic determinants simultaneously. One of the exceptions concerned the large-scale association study in 10,601 subjects by Fredriksen et al.[3] which described the association between thirteen DNA variants relevant to one-carbon metabolism and plasma tHcy. Statistically significant associations with plasma tHcy were found for *MTHFR*667C>T, *MTHFR*1298A>C, *MTR*2756A>G, and the *CBS*844_845ins68 variant and not for *MTRR*66A>G, *MTHFD*1958G>A, *BHMT*716G>A, *CBS*699C>T, *TCN2*67A>G, *TCN2*776C>G, *SLC19A1*80G>A,

*PON1*163T>A and *PON1*575A>G. Furthermore, the genetic associations for eleven additional one-carbon metabolites were described[3]. Several of the 179 identified studies also measured additional one-carbon metabolism intermediates, mostly folate and cobalamin levels, generally to evaluate gene-environment interaction on plasma tHcy. Only very few studies reported on multi-locus effects for the small number of DNA variants that were simultaneously measured. Indications for presence of interactions between DNA variants were reported by some. For instance, Devlin et al.[30] described the marginal and interaction effects for four polymorphisms and reported a plasma tHcy increasing interactive effect for *MTHFR*677TT and *SLC19A1*80GG genotypes. An interaction for *TCN2*776C>G and *MTRR*66A>G genotypes on plasma tHcy was reported by Aléssio et al.[42] while Yang et al.[7] described an interaction between *MTHFR*677C>T and *MTRR*66A>G on plasma tHcy.

Overall, research efforts by numerous groups in recent years mainly focused on evaluating previously identified nonsynonymous variants in genes related to remethylation of homocysteine to methionine and folate cycle. Association to disease phenotypes and plasma tHcy were mostly studied using less than 400 subjects. Relatively few studies have concentrated on variation in one-carbon metabolism genes related to transsulfuration and transmethylation. Only three studies, including our study described in chapter 2, specifically aimed at identification of new or established genetic determinants of plasma tHcy using a non-hypothesis-driven genome-wide linkage approach[53,58,59]. Contradicting (linkage and) association results among studies, which may be due to population heterogeneity or chance, were found for many DNA variants, also for those that are known to affect the product encoded by the gene. So after years of study, the *MTHFR*677C>T variant still stands out as most important and only established marginal genetic determinant of plasma tHcy concentrations. Other potential determinants, like *CBS* 31 bp VNTR, have been identified but await convincing replication in other populations.

Our genome-wide linkage study suggested positional candidate regions and showed no overlap with the two other genome-wide linkage studies for plasma tHcy

Three genome-wide linkage studies for plasma tHcy, including our study described in **chapter 2** of this thesis, have been published in 2005 and 2006[53,59]. The heritability estimate of 44% for plasma tHcy in our study indicated potential for the identification of genetic determinants for plasma tHcy in our study population. However, we were not able to identify QTLs with convincing statistical evidence. Also, we could not replicate the linkage signals for three regions that were found by Souto and colleagues[53] based on 21 extended pedigrees. In this Spanish linkage study a new candidate gene for plasma tHcy, nicotinamide *N*-methyltransferase (*NNMT*), was revealed[53]. However, subsequent association studies for *NNMT* variation in other populations could not

confirm a strong role for these variants in plasma tHcy concentrations[35,54]. The third linkage study that was published shortly after ours by Kullo et al.[59] identified four linkage regions with statistically significant and tentative evidence in 390 non-Hispanic white sibships; the regions did not contain any known genes involved in homocysteine metabolism. Again, no overlap in linked regions for this study, that of Souto and colleagues[53] and ours was found. Differences between the three linkage studies may be ascribed to genetic and/or environmental heterogeneity among the populations or to the relatively low power of the studies. We tried to replicate our findings in the Homocysteine in Families (HOFAM) study, a Dutch family study of patients with vascular disease and hyperhomocysteinemia comprising 306 individuals in 51 pedigrees[60]. Three markers located in the linked region of chromosome 16 and one marker located in the linked regions on chromosome 12 and 13 were genotyped. A non-significant multipoint LOD score of 1.15 was found on chromosome 16 for age, sex and *MTHFR*677C>T adjusted plasma tHcy; no other signs of linkage were found (data not published). We conclude that we contributed to additional positional candidate regions that may harbour QTLs for plasma tHcy regions.

Our extensive candidate-gene association studies supported MTHFR677C>T and variants in CBS as main but modest marginal determinants of plasma tHcy and indicated new candidates

To our knowledge, **chapters 3 and 4** in this thesis described the most extensive candidate-gene association studies for fasting as well as post-load plasma tHcy in terms of number of measured DNA variants in a single study population, so far. The results of our association study of 79 DNA variants in 40 one-carbon metabolism-related genes in **chapter 3** underlined the major importance of the DNA variants *MTHFR*677C>T, *CBS* g.14037(31bp)16-21, and *CBS*844_845ins(68bp) in our population as published by us previously[45,61-63]. The latter two variants showed strong LD. Other DNA variants that had not been studied in relation to one-carbon metabolites by others were identified, although not with statistically significant evidence (e.g. rs4819205 in *FTCD*, a gene encoding formiminotransferase cyclodeaminase and involved in the folate cycle). The set-based analysis of sub-pathways emphasized the major role for genetic variation in transsulfuration genes in plasma tHcy levels, especially post-load. We also found association to genetic variants of methyltransferase genes (e.g. SNPs in *NNMT*, *DNMT3A*, *AHCY*) in line with reports from Souto et al.[53] on *NNMT* and Gellekink et al.[49] on catechyl-o-methyltransferase (*COMT*). The strong association between three SNPs in ATIC and plasma tHcy was disputed based on further study by our group[38]. All in all, the application of a large-scale custom SNP-array approach identified new potential candidate genes that may harbour variants with somewhat smaller influence on plasma tHcy compared to *MTHFR*677C>T in the general population. Our study

provided new leads for the future search for genetic determinants of plasma tHcy and one-carbon metabolism in general.

Trait-specific and shared genetic determinants of one-carbon intermediates illustrated

The measurements of folate, methionine as well as homocysteine concentrations in **chapter 3** illustrated some similarity in underlying genetic risk factors but mostly dissimilarity in main single locus effects for these three one-carbon metabolism intermediates that have all been associated to disease. The *MTHFR*677C>T was a main determinant of serum folate as well as a strong determinant of plasma tHcy. This finding has also been reported by others (e.g. [3,64]). All three metabolites showed association to variation in *CUBN* although to different SNPs. For other DNA variants and genes dissimilarity in associations with the three metabolites was observed (e.g. plasma tHcy was strongly influenced by DNA variants in *CBS* in contrast to methionine and folate concentrations, only methionine was strongly negatively associated to DNA variants in *BHMT*). The large-scale study by Fredriksen et al.[3] described above also found differences and overlap among genetic determinants of several one-carbon metabolites. The importance of awareness of multiple effects of genetic variants in the interpretation of Mendelian randomization studies has already been emphasized; the association between the genetic variant that predisposes to a certain environmental phenotype and disease that is studied via Mendelian randomization may be confounded by pleiotropic effects[65]. Hence, knowledge on the metabolic profile of DNA variants facilitates application and interpretation of Mendelian randomization studies and elucidation of pathogenic mechanisms.

Multi-locus analysis: a combination of five genetic variants explained 17% of variation in plasma tHcy concentration

In **chapter 4** we performed a multi-locus analysis including haplotype association and logic regression analysis for 14 candidate DNA variants. Previous analyses of multi-locus genotype effects have focussed on first-order statistical interactions for a limited number of DNA variants using traditional regression or ANOVA techniques[11,30,38,42,66] or evaluated interaction for *MTHFR*677C>T with the total number of mutant alleles in other DNA variants[67]. Our choice for logic regression analysis allowed a thorough search over 14 DNA variants for multi-locus effects that may include higher-order interactions. This allowed the identification of a relatively common 5-locus genotype that was highly associated with increased fasting plasma tHcy: 10% of subjects that carried the genotype {*CBS* 31bp VNTR 18-18 and *FOLH1*1561CC and (*MTHFD*2011GG or *BHMT*595GA) and *MTHFR*677CT/TT} had on average 5 µmol/L higher plasma tHcy concentrations compared to the 90% that did not. This multi-locus genotype included

pair-wise statistical interactions that had been identified by us (*MTHFR*677C>T and *CBS* 31 bp VNTR) previously[68]. By the way, this multi-locus genotype showed a much weaker but still significant association to serum folate concentrations and explained 6% of its variation. Our study emphasized that it is the combination of multiple detrimental genetic variants that has the potential to explain observed variation in plasma tHcy, and hence the polygenic and interactive qualities of this complex trait.

It also indicates that the use of multi-locus genotypes in Mendelian randomization studies can achieve higher power compared to studies that rely on the much smaller single locus effect of *MTHFR*677C>T. For example, homozygous mutant state for this genotype results in on average 2 µmol/L higher plasma tHcy compared to CC and CT genotypes and can be found in the population with a frequency of ~12%. The multi-locus genotype, with similar frequency in the population, resulted in 5 µmol/L higher plasma tHcy. Assuming that a 1 µmol/l difference in plasma tHcy is related to 8% risk reduction for venous thrombosis (VT), one would expect to find 20% and 40% increased VT risk for the *MTHFR*677TT and high homocysteine multi-locus genotype, respectively. Evaluation of these effects in a case-control setting with 80% power and an α of 5% would require 4107 cases and controls for *MTHFR*677C>T and 1142 cases and controls for the multi-locus genotype. This potential relaxation of sample size requirements when using multi-locus genotypes that produce substantial differences in intermediate phenotypes has been touched upon in one of the recent papers by Ebrahim and Davey Smith[69]; the evaluation of associations between *combinations* of unlinked polymorphisms and disease has been termed 'factorial Mendelian randomization'.

Identification of new candidate genes for NTD that are not dependent on plasma tHcy

The many studies in recent years into the role of genetic variation in NTD aetiology have focused on genes related to homocysteine and folate metabolism, including *MTHFR*, *MTHFD*, *MTR*, *MTRR* and *SLC19A1*. *MTHFR*677C>T and *MTRR*66A>G are considered genetic risk factors for NTD but a large part of the genetic contribution to NTDs still remains to be identified[70]. **Chapter 5** described the most extensive candidate-gene association study for NTD to date. The application of a custom SNP-microarray approach in a case-control setting supported involvement of *SLC19A1*80G>A and identified new potential candidate DNA variants in the aetiology of NTD, including SNPs in *CUBN* and *TRDMT1*. These associations have not been investigated by others to our knowledge.

Previous studies by us and other research groups have indicated the potential role for impaired one-carbon-related methylation in NTD aetiology[71]. Our candidate gene study described in **chapter 3** pointed towards a role for *DNMT1*, encoding a DNA methyltransferase, in folate concentrations but variation in this gene was not found

among the strong determinants of spina bifida risk. However, a SNP in *TRDMT1* (previously known as *DNMT2*) was associated with decreased risk for spina bifida as well as increased folate concentrations. TRDMT1 does not methylate DNA but aspartic acid transfer RNA (tRNA(Asp)). In addition, our study suggested involvement of uptake of vitamin B_{12} and folate in NTD aetiology; SNP variation in *CUBN* and *SLC19A1* was associated to spina bifida risk; however, only the first showed association to RBC folate and cobalamin concentrations.

Neither of the two DNA variants that showed strong (negative) association to spina bifida risk showed nominal association to plasma tHcy, but we did see decreased plasma tHcy in line with expectations. Also, spina bifida risk was not strongly affected by *MTHFR*677C>T and *CBS*844_845ins(68bp) genotypes. Hence, our findings do not support a key role for plasma tHcy in NTD causality. Instead, the simultaneous analysis of one-carbon metabolite concentrations and spina bifida-affected individuals underlined the potential importance of methylation processes and cobalamin concentrations in NTD aetiology, as proposed by others[71-73].

UCP2 involved in venous thrombosis aetiology and plasma tHcy: what is the mechanism?

We were the first and only group to investigate the *UCP2* 45 bp del/ins for its association to plasma tHcy and involvement in recurrent venous thrombosis (RVT) risk. This study, described in **chapter 6**, was triggered by the association between plasma tHcy and RVT[74], the fact that oxidative stress has been proposed to underlie the prothrombotic actions of homocysteine[75], the potential role of UCP2 in oxidative stress regulation[76] and our initial findings of an association between this variant and plasma tHcy in a small study sample (data not published). Also, some studies indicated involvement of UCP2 in development of atherosclerosis[77] and cardiovascular risk[78]. The results of our study in **chapter 6** indicated that *UCP2* 45bp del/ins may moderately increase post-load plasma tHcy as well as RVT risk. The small effect of the genetic variant on plasma tHcy (<6% change in plasma tHcy concentration versus wild-type) and the relatively large effect on RVT risk (OR 1.4) suggests that the genetic variants exerts at least a part of its effect via other ways than plasma tHcy concentrations. In addition, it suggests that the high levels of plasma tHcy in RVT cases should be largely due to other causes. This is supported by the fact that logistic regression analysis for RVT including both the *UCP2* 45bp del/ins and plasma tHcy concentrations as independent variables showed an OR comparable to the crude ORs for *UCP2* 45bp del/ins genotypes and a very strong association for plasma tHcy and RVT. Genetic association analysis stratified for plasma tHcy quartiles, hampered by low power, does not show indications for effect modification (data not shown). Hence, the underlying mechanism linking UCP2 and homocysteine to RVT is still unclear and elucidation relies on future studies.

New hybrid design offers powerful, valid, and flexible alternative for studies into maternal and offspring genetic effects

The comparison of the new hybrid design with four existing designs that all allow the evaluation of offspring and maternal effects as described in **chapter 7** has increased insight into their relative performance in terms of power and underlying assumptions. The study is an addition to existing methodological papers in this research area[79-82] and a continuation of the work by Weinberg and colleagues[83,84]. In line with these papers, we illustrated that power for the popular case-parent triad and control-mother/case-mother dyad designs were more or less similar depending on the underlying genetic model, though different underlying assumptions have to be met, and we, again, showed the surplus value of hybrid designs in terms of power, validity, and flexibility. In addition, we offered an alternative hybrid design for a situation in which fathers of controls may be hard to recruit or control-mother dyads are readily available. We have also explicitly laid out the general case-triad/control-triad design of which all the illustrated designs are subsets of. Our research described in **chapter 7** also set off a study into the optimization of the statistical analysis of case-mother/control-mother dyad data that has been recently published[85]. Both publications have extended the toolbox for genetic epidemiological research into offspring and maternal genotype effects that may be especially valuable in the field of complex congenital and early-onset disorders, like NTDs. We await application of the new design and techniques of analysis in the future.

Evaluation of four multi-locus techniques of analysis using simulated data may guide application to real data

Logistic regression, MDR, logic regression and set-association are popular methods of analysis that are applied nowadays (e.g. [86-89]) and are likely to be applied in future studies to identify multi-locus effects. The MDR, logic regression, and the SumStat set-association methods have been described and evaluated by their developers using simulated and real data prior to or shortly after the introduction of the methods to the scientific public[90-95]. However, the number of different underlying statistical interaction models for which the methods were evaluated was restricted and the methods were not simultaneously applied to the same data. The application of logistic regression, MDR, logic regression and set-association analysis to simulated data based on six different two-locus causal genetic models was described in **chapter 8**. The results confirmed that the methods were performing well in general, showed that logistic regression performed worse and that small differences were observed for the other methods depending on underlying models. They also demonstrated that all methods were capable of identifying interacting causal DNA variants in the absence of single-

locus marginal effects, but also showed that the ability to do so was impaired in the presence of interaction and genetic heterogeneity (i.e. model in which the presence of one out of two two-locus genotypes without single marginal effects increased disease risk). The latter was touched upon by Ritchie et al.[91] previously and an adaptation of the MDR method to deal with genetic heterogeneity has been proposed recently[96]. In addition, we have been the first to demonstrate the surplus value of incorporation of interaction terms in the set-association method. Several studies reporting the simultaneous application of traditional logistic regression, set-association, logic regression, or MDR analysis to real multi-locus data have been published after our study[87,88,97,98]. These all showed that the different methods produced overlapping results but also that some discrepancies exist. The development and improvement of the MDR and logic regression methods has been ongoing[96, 99-104] and new multi-locus methods are being introduced (e.g. [105]). Developmental changes to the core procedures of the evaluated methods of analysis will render our results outdated at some point in time. For now, the results of our study may be used to guide and facilitate the analysis and interpretation of multi-locus data as for our studies described in **chapters 4 and 5** of this thesis.

9.2 Strengths and weaknesses of our studies

9.2.1 Study designs and utilized study populations

Modest sample sizes inhibited identification of genetic determinants of plasma tHcy and NTD

The linkage analysis performed in chapter 2 allowed hypothesis-free scanning of the genome for plasma tHcy QTLs without the need to predefine candidate genes. The linkage design is not sensitive for population stratification and it can deal with allelic heterogeneity. However, it is not a powerful approach in complex traits in which each locus is expected to contribute only a small percentage to trait variance[106]. Indeed, our power calculations showed that our study, that included 264 individuals in 13 extended pedigrees, had 75% power to identify a QTL that explains 25% of trait variance with a LOD score ≥ 1. We were thus underpowered to find loci that explain, say, ≤ 4% of plasma tHcy variation. The completion of the Human Genome Project[107], the International Haplotype Mapping (HapMap) project[108,109], and the advances in genotyping technologies have enabled a powerful equivalent of genome-wide linkage scans for complex diseases: genome-wide association (GWA) or linkage disequilibrium (LD) scans. Commercial SNP-arrays that are capable of tagging most of the common variation in the human genome have been introduced[109] and examples of successful applications have emerged over the last few years[110-112].

All studies described in chapters 3 through 6 relied on evaluation of candidate DNA variants in a population-based association approach to identify genetic risk factors. Population-based association studies are popular in complex diseases for their relatively high power compared to linkage and family-based association designs. However, the search for DNA variants with small marginal contributions to a continuous trait or disease susceptibility in the study population still requires a large number of individuals to reach sufficient precision in estimating and testing genetic effects[113]. The case-control studies described in chapters 5 and 6 were based on 386 controls and 161 cases, and 190 controls and 180 cases, respectively. Our study in chapter 5 was only slightly underpowered (based on alpha 0.05 and observed MAF and odds ratios). The case-control study for spina bifida risk was underpowered to detect DNA variants with small contributions to disease risk; adjustment for multiple comparisons further diminished the power of this study. Also, the sample sizes in chapters 3 and 4 were too small for estimation of small genetic effects with high precision (e.g. 200 individuals, DNA variant coefficient of determination R^2 = 2%; α 0.05; power 52%).

In chapter 5 we aimed to identify genetic determinants that are involved in the aetiology of spina bifida. The measurement of DNA variants in spina bifida cases and controls did allow us to do so. However, this design does not permit differentiation between offspring and maternally-mediated genetic effects and has reduced power to identify maternal genetic risk factors for spina bifida (see chapter 7).

Well-phenotyped and homogeneous study samples enhanced identification of genetic risk factors

Population-based genetic association studies may suffer from bias due to the existence of population stratification. In order to minimize the chance for population stratification bias, we have excluded the non-Caucasian individuals (based on questionnaire data) that were present in the population-based samples of individuals that served as study or control population in chapters 3 through 6. The spina bifida cases and recurrent venous thrombosis patients described in chapters 5 and 6 were all of Dutch descent (based on surname information of patients and (grand)parents) and the latter were obtained from the same geographical area as the control samples. Several studies based on real and simulated data have shown that false positive and biased associations due to population stratification are likely to be limited in studies of moderate size, except for some unrealistic scenarios[114,115,116]. Hence, the presence of residual confounding due to population stratification is expected to be limited for our studies. The bias resulting from population stratification can increase substantially, however, when sample size increases[117].

Even though confounding due to other factors than population stratification in genetic association studies is unlikely, it may still occur by chance, especially in case of

small study populations and genetic variants with low MAF; the assumed similar distribution of outcome-associated factors over the genotype groups under study may not hold in these cases. In the genetic association studies for plasma tHcy concentrations, we have performed crude analyses as well as analyses in which plasma tHcy was adjusted for age, sex, and, in chapter 4, creatinine, important determinants of plasma tHcy. The adjustment for these determinants of plasma tHcy may also have increased power due to the removal of variation in outcome attributable to these factors.

Causal heterogeneity is a characteristic of complex diseases. Selection of patients that are homogenous with regard to their clinical phenotype may decrease causal heterogeneity in the study population and, hence, increase the power of the study. Several studies have suggested differences in genetic susceptibility factors for the different types of NTDs[118-120]. The spina bifida patients that have been studied in chapter 5 were selected from an NTD patient collection that has been ascertained at different time-points via a Dutch society for patients with central nervous system defects and via the paediatric neurology department of the Radboud University Nijmegen Medical Centre. For our study, we included only those patients with sporadic spina bifida aperta, an open closure defect located at the caudal part of the neural tube. The venous thrombosis cases described in chapter 6 concerned only those patients that experienced at least 2 episodes of venous thrombosis. Venous thrombosis mainly manifests itself in two related conditions: deep vein thrombosis (DVT) and pulmonary embolism (PE). Around 90% of PE are the result of thrombi in the deep veins of the lower extremeties. Although Factor V Leiden mutation is a strong risk factor for DVT and not PE, in general the two conditions are viewed as a single disorder with shared risk factors[121].

9.2.2 Measurements of main variables

Availability of standardized measurements of fasting and post-methionine load plasma tHcy concentrations enhanced identification of genetic risk factors

All blood samples in the control population and RVT cases were collected after an overnight fast to minimize the inter-individual variation due to short-term dietary intake. In addition, plasma tHcy measurements 6 hours after a standardized oral methionine-load were determined. All studies relied on a single measurement of plasma tHcy. The standardized performance of multiple measurements may reduce the analytical variation and thus the intra-person variation (which is the sum of the intra-person biological and analytical variation (also called coefficient of variation (CV))) and, hence, the power of the study. The intra-person variability for plasma tHcy is ~8% (6.5% using the method used in our studies[122]) and a single measurement in cross-sectional studies in healthy populations, although it may underestimate the effect under study, is generally regarded as sufficient[123].

Post-load and fasting plasma tHcy concentrations show strong correlations (see chapter 3) and share underlying risk factors. The stress induced by a high oral load of methionine has been used in the past to identify genetic defects in the vitamin B6-dependent transsulfuration pathway[123]. Since then, post-load plasma tHcy has been associated with cardiovascular disease risk[124], also independent of fasting plasma tHcy[125]. Our group found a considerably higher heritability for post-load concentrations compared to fasting plasma tHcy[60]. Hence, the measurement of post-load plasma tHcy as well as fasting plasma tHcy may have increased our chance of identifying genetic determinants involved in plasma tHcy-related disease traits.

Studies did not fully utilize increase in knowledge about DNA variation and advances in genotyping technologies in recent years

The association study described in chapter 4 is based on low-throughput genotyping of DNA variants, performed between 1995 and 2007. Generally, only one DNA variant in a gene was measured. In addition to SNPs, fine-scale structural variants including insertions, deletions, and VNTR variants were measured and analyzed. It has become clear that the human genome entails many small, intermediate, and large-scale structural variations, including copy-number variations (CNVs)[126] that may show an important influence on disease traits[127,128]. We haven't included the latter types of variants in our studies; insight in their prevalence in the human genome, adequate genotyping techniques, and methods of analysis, are still under development[127,129].

The selection of genes and DNA variants in the study described in chapters 4 and 6 was based on gene sequencing of coding regions (in patient populations) and reported findings from others. DNA variants with proven or potential functionality (i.e. affecting the amount or activity of the encoded protein) were prioritized. The 50 genes selected for the studies described in chapters 3 and 5 were involved in different folate-related processes. The DNA variants were selected from literature and databases during 2002. Again, variants that changed the function or regulation of a gene product and those that had already shown association with a disease trait were prioritized. If not available, potentially functional, confirmed SNPs or SNPs within or near exons and in the promoter region that may be in LD with unknown causal variants were selected. This approach increased our prior chance of identification of biologically meaningful DNA variants for plasma tHcy and NTD in population studies, reduced the amount of genotyping efforts and number of tests that needed to be performed, and fully exploited the extensive knowledge on homocysteine and one-carbon metabolism and our current understanding of NTD aetiology[130]. It did however not lead to coverage of all variation present in the selected genes and pathways in the general population. Given that there is incomplete understanding of the molecular pathways underlying disease traits and that the full function of the studied genes and their dependence on

DNA variation is not known, this may have resulted in incomplete assessment of the association between the genes and the traits of interest.

The selection of SNPs with the aim of gene or regional coverage of SNP variation has been improved dramatically by revelation of the sequence of the human genome and the execution of the HapMap Project[108,109]. It is now possible to efficiently select a subset of SNPs that capture (or tag) the common SNP variation that is present in a certain genomic region or even the whole genome in the HapMap populations prior to any genotyping effort. In addition, the improvements in genotyping technologies have reduced the costs of large-scale genotyping, increasing the feasibility of extensive candidate-gene or GWA studies in large study populations. Indirect measurement of rare variants (MAF<0.01) via a tagging approach remains difficult however, due to poor capability of common SNPs to tag rare variants.

The genotyping of the DNA variants analyzed in the studies presented in chapters 4 and 6 was performed at different time-points over a period of 12 years and based on low-throughput genotyping techniques. Unfortunately, genotyping efforts have not been completed for all ascertained individuals due to low DNA quality and absence of DNA. Applied quality-control measures to detect and minimize genotyping errors varied per genotyping effort and included duplicate genotype assignment, duplicate measurements, blank control wells, and tests for HWE. All analyzed DNA variants in chapters 4 and 6 complied to HWE, with exception of the multi-allelic 31 bp VNTR in *CBS* which showed only a slight deviation (P-value $_{HWE}$ 0.022). Large deviations from HWE have been used as an indication for genotyping errors, although some discourage the use of HWE to track genotyping errors[131,132]. Hence, tentative interpretation of allelic association results for the 31 bp VNTR in *CBS* may be warranted but does not necessitate omission of the results.

Genotyping for the studies in chapters 3 and 5 was mainly performed using a customized SNP-micro-array approach developed and executed by Asper Biotech, Estonia, from 2003 until 2005. Additional non-SNP variants were genotyped separately. Of the 154 DNA variants that were selected, only 79 and 87 variants in chapter 3 and 6, respectively, passed the quality controls. These included a MAF>0.05 or 0.02, <25% missing genotype data, and HWE P-value>0.01, in order to ensure exclusion of genotyping errors and those DNA variants for which we were largely underpowered in our studies. We chose for the lenient percentage of 25% missing genotype data to be able to include variants for which genotyping had only been performed in a subset of the data and to exclude DNA variants with low success rates indicating low assay quality. The high failure rate was largely due to SNPs that were non-polymorphic in our population but also to failure of HWE. Errors in design and execution of the SNP micro-array approach have contributed to this inability of genotyping some of the selected SNPs. The extensive experience with SNP-arrays that has been acquired by commercial companies and research groups in the last 5 years has dramatically improved quality and genotyping success rates for current SNP-array platforms.

9.2.3 Statistical analysis

Use of advanced analyses optimized evaluation of DNA variant associations but opportunities for improvement exist

The advantages of haplotype analysis have been discussed extensively by others (e.g.[139]). We used Whap[133] to perform haplotype estimation and analysis. With help of Whap a weighted regression analysis was performed that takes the potential uncertainty in the deduced haplotypes for an individual into account and, hence, prevents the invalid results that may arise in the application of methods that use the most likely haplotype pair per subject[134,135].

The multi-locus logic regression analysis allowed the identification of a common multi-locus genotype that was highly associated to plasma tHcy concentrations. Simulation studies described in chapter 8 and studies by Kooperberg et al.[92] and Ruczinski et al.[93] showed the ability of this technique to identify interacting susceptibility loci. The application of a simulated annealing algorithm as implemented in logic regression to identify the best model for available genetic variants and outcome trait may result in overfitting of the model on the data with limited fit to external data. Hence, we used cross-validation techniques to choose the best model size. However, several models of different sizes performed similar in the cross-validation analysis. Increased sample sizes will improve model selection and follow-up studies in independent replication samples will learn whether the selected model and multi-locus genotype is applicable to other populations. If the goal is to evaluate the overall importance of the analysed SNPs for trait outcome instead of identifying a single best model, a Monte Carlo variant of logic regression analysis can be applied[103].

The logic regression analysis described in chapter 4 was hampered by the presence of a large number of samples with incomplete genotypes for the analyzed DNA variants. Also, we had a high drop-out rate for SNPs selected for the SNP-microarray approach. In addition to loss of power, the presence of missing values and complete-case analysis may result in biased effect estimates, depending on the underlying pattern of missing values. The precision and validity of the effect estimate can be improved by application of (multiple) imputation techniques[136]. Several methods to impute missing genotypes have been described and evaluated[137]. Successful application of these methods in the studies described in this thesis will however be impaired by the low SNP density and LD.

Chapter 3 reports the genetic association between the DNA variants and three one-carbon intermediates separately. This allowed a global evaluation of the similarities and dissimilarites in associated DNA variants across traits. Also, chapters 5 and 6 measured genetic association for multiple disease phenotypes. The individual analysis of the separate traits did however increase the multiple testing burden and did not

exploit the existing correlations between the phenotypes. More advanced methods for the simultaneous genetic association analysis for multiple correlated traits have recently been proposed[140,141]. Such methods may lead to more powerful approaches in gene identification and increase insight in genetic underpinnings of disease traits.

9.3 Future perspectives

Inconclusive results RCTs for plasma tHcy-reducing therapy: still analyse genetic determinants plasma tHcy

In the last decades, convincing evidence for causal involvement of homocysteine concentrations in several chronic diseases has been described: homocystinuric patients exhibit vascular disease, an abundance of retrospective and prospective observational studies confirm a positive linear relation between plasma tHcy concentrations and risk for several disease phenotypes, meta-analyses of Mendelian randomization studies for vascular disease and NTD, of which most have focussed on *MTHFR*677C>T, have shown overall positive results, periconceptional folic acid supplementation reduces the risk of NTD pregnancies, B-vitamin therapy has been shown to reduce cardiovascular disease risk in homocystinuric patients, and plausible mechanisms for pathogenic effects of plasma tHcy have been reported. All these findings support a causal and modifiable relation between homocysteine concentrations and disease.

Evidence from randomized intervention trials with plasma tHcy-lowering therapy was eagerly awaited to conclusively underscore this 'homocysteine hypothesis' for vascular disease. However, unconvincing effects of B-vitamin therapy on vascular disease risk and mortality[142,143], with an exception for ischemic stroke[144], were reported. Also, a decreased risk for (recurrent) venous thrombosis after homocysteine-lowering therapy could not be convincingly confirmed[145,146]. And although one meta-analysis of randomized clinical trials underlined the positive effects of B-vitamin therapy on flow-mediated dilatation (FMD), a marker of endothelial function[147], a recent meta-analysis by Potter et al.[148] provided negative results for FMD and carotid intima-media thickness after long-term homocysteine-lowering treatment with B-vitamins in subjects with history of stroke. Overall, the trial results for vascular disease were disappointing and led to increased speculation about non-causality for plasma tHcy in vascular disease causation[149]. Several studies have however underlined that some characteristics of the randomized clinical trials may have impaired the ability to convincingly show the positive effect of B-vitamin therapy. These include insufficient sample sizes to reliably detect the effects of B-vitamin therapy, especially in populations where folic acid fortification has been introduced, additional deleterious effects of B-vitamin therapy, a too short duration of B-vitamin therapy to reverse the negative effect of exposure

to high homocysteine levels in the past, and the fact that only specific groups may be susceptible for positive influence of B-vitamin therapy[142,150,151]. Evidence from ongoing studies may contribute to the elucidation of the apparent inconsistencies among the trials and between the observational and experimental studies.

As already described in chapter 1, knowledge on the genetic aetiology of plasma tHcy concentrations that is generated by epidemiological studies can contribute to the elucidation of the mechanistic role of plasma tHcy and aetiology of disease phenotypes. However, studies by several research groups into genetic determinants of plasma tHcy in the past years have unfortunately led to relatively little progress in knowledge on genetic factors for plasma tHcy (see section 9.1); the *MTHFR*677C>T is still the most replicated and important determinant of plasma tHcy concentration but with a modest marginal effect that is heterogeneous among different populations. As a result, Mendelian randomization studies have been hampered by the unavailability of DNA variants with strong, robust, and specific effects on plasma tHcy concentration. One of our major findings described in this thesis was the identification of a common multi-locus genotype that predisposed to extremely diverse concentrations of plasma tHcy concentration. These findings may facilitate more powerful Mendelian randomization studies. Future studies will need to show whether the large effect of the common multi-locus genotype identified in our population in chapter 4 can be replicated in other populations. They can also focus on identification of plasma tHcy associated multi-locus genotypes based on a larger and/or diverse number of DNA variants. Then, studies into association between the multi-locus genotypes and disease traits may be performed more efficiently.

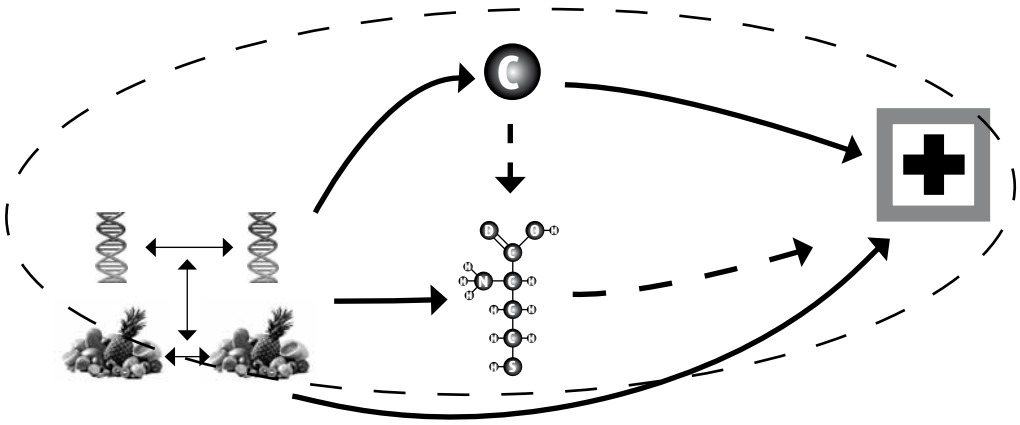

Figure 9.1 *Simplified presentation that depicts the potential intermediate role of one-carbon metabolism (OCM) biochemical parameters between fundamental genetic and environmental risk factors and the ultimate disease phenotype.*

But also study additional one-carbon metabolism biochemical parameters ánd disease phenotypes

The ambiguity about causality and modifiability of plasma tHcy in common disease aetiology and the established importance of one-carbon metabolism in several essential physiologic processes and diseases argue for measurement of a wider variety of one-carbon metabolites (e.g. folate, methionine, S-adenosylhomocysteine, asymmetric dimethylarginine, B-vitamin concentrations). Preferably, markers that more specifically reflect essential one-carbon metabolism related processes are included as well. For instance, markers of DNA methylation potential (e.g. determination of S-adenosylhomocysteine/S-adenosylmethionine ratio (see also[152]), global and/or gene-specific DNA methylation), nucleotide synthesis (e.g. determination of uracil incorporation in DNA building blocks), oxidative stress (e.g. glutathionine, malondialdehyde), and endothelial dysfunction (e.g. flow-mediated dilatation) can be measured. The genetic analysis of this wider variety of one-carbon metabolism biochemical parameters in conjunction with plasma tHcy and disease facilitates the identification of new intermediates and increases insight in one-carbon metabolism-related disease mechanisms (see Figure 9.1). It requires the simultaneous measurement of multiple one-carbon metabolism biochemical parameters in large numbers of samples. The latter is facilitated by the recent development of advanced analytical platforms that are focused on one-carbon metabolism[153] and improvement in metabolite measurement in general.

This principle of measurement of multiple metabolites and disease was demonstrated by our simultaneous study of folate-related polymorphisms for spina bifida as well as folate, plasma tHcy, and cobalamin concentrations. The two strongest genetic predictors of spina bifida risk, an intronic SNP in *CUBN*, involved in uptake of cobalamin, and an intronic SNP in *TRDMT1*, an RNA methyltransferase, also showed association to folate and/or cobalamin concentrations which underlined the potential importance of these genes and directly provided clues about a possible disease mechanism. These two genes have not been evaluated for involvement in spina bifida aetiology by others, to our knowledge. Replication of associations of these genes to spina bifida risk and folate and cobalamin in other independent populations will be a next step in characterization of the role of these genes in spina bifida aetiology. In addition, we showed that the link between variation in *UCP2*, an oxidative stress-related gene, and recurrent venous thrombosis can be identified in human population studies. How plasma tHcy is positioned in this disease mechanism remains to be elucidated. Additional studies with measurements of markers of oxidative stress will provide valuable information.

Future genetic epidemiological studies into plasma tHcy and related one-carbon metabolism biochemical parameters and diseases will need to exploit developments in the field of genetic epidemiology

As stated previously, numerous efforts of multiple research groups to elucidate the genetic underpinnings of plasma tHcy via genetic epidemiologic studies have been relatively unsuccessful in the past years. The disappointing results are related to the application of designs and analyses that do not face up to the complexity of the trait under study, even though its complexity is expected to be less than that of the heterogeneous disease trait itself. Challenges for future genetic epidemiologic studies include the up-scaling of sample sizes, the performance of extensive genotypic characterization, the utilization of large-scale sequence data, the incorporation of epigenetic and gene expression data, and application of advanced study designs and analyses. Overcoming these challenges allows for a more efficient evaluation of the role of genome variation and may spur the elucidation of one-carbon metabolism involvement in disease in coming years. A brief explanation of each challenge is given below.

Up-scaling of sample sizes

High precision estimation of the small effects of a large number of DNA variants and their joint interactions on a number of quantitative disease traits and, especially, qualitative disease outcomes requires large samples sizes (>>1000)[154]. Also, availability of independent replication samples is needed to evaluate whether the identified effects can be confirmed in other populations. One must bear in mind however, that genuine heterogeneity between study populations (e.g. variability in LD pattern, variability in exposure to environmental factors) may cause non-replicability[155]. The necessary increase in sample size is facilitated by the initiation and existence of biobanks and international collaborations (consortia) like the Wellcome Trust Case Control Consortium (WTCCC)[156] and POLYGENE[112]. As already touched upon in section 9.1, several researchers have used available samples from large-scale studies and consortia to powerfully study the associations between one-carbon metabolites, disease, and DNA variants ('Norwegian Colorectal Cancer Prevention (NORCCAP)' study[2,3]; 'Hordaland Homocysteine study'[4]; 'Cardiovascular Risk in Young Finns Study'[5]; 'Health in Men study'[6]; 'National Health and Nutrition Examination Survey DNA Bank (NHANES) III'[7];'Women's Health Study'[8]; MEGA study[157]). Also, recent genome-wide association initiatives contain valuable data on genotype and one-carbon metabolism-related phenotype data (see below). New studies should exploit these existing and future large-size sample collections and collaborations in the search for and characterization of genetic determinants of one-carbon metabolism phenotypes and related diseases. Also, standardization and harmonization of blood sample collection, metabolite measurements, and collection of environmental information in population-based cohorts is

required to allow for evaluation of gene-environment interaction and deal with population heterogeneity[158].

Extensive genotypic characterization

As stated previously, the availability of high-density data on genetic variation in a number of populations allows the efficient selection of tagSNPs (in addition to functional DNA variants) for common variation in candidate genes or genomic regions of interest. It has also led to the development of genome-wide SNP-arrays that allow a relatively powerful scan of the human genome for trait susceptibility loci. Results of a genome-wide association study (GWAS) for plasma tHcy concentrations are expected in due time (personal communication van Meurs[159]). GWAS efforts may show whether the genome contains previously unknown DNA variants that show robust single or multi-locus association to plasma tHcy; these may be of more influence than the candidate loci that have been studied so far. Two separate studies have recently reported on single locus GWA results for methionine[160] and cobalamin concentrations[161]. No significant associations for methionine were reported but strong associations between SNP variants in fucosyltransferase 2 (*FUT2*) and plasma cobalamin in 1,658 subjects and an independent replication sample (n = 1,059) were found; *FUT2*, that has not been previously considered as a candidate gene for cobalamin concentrations, is proposed to affect cobalamin absorption and associated genotypes showed ~8% difference in plasma vitamin B_{12} levels for the different genotype groups. The SNP also showed GW association, although much weaker, to plasma tHcy concentration[161].

In addition, genome-wide copy-number variation (CNV) measurements using specific probes positioned on SNP-arrays and measurement of CNVs in large populations is an area of attention and development. CNVs have been shown to influence several phenotypic traits independent from SNPs[127] and may well contribute to variation in one-carbon metabolism phenotypes.

Commercial SNP-arrays containing 12[162] and 57[163] functional and key polymorphisms in candidate genes involved in one-carbon metabolism are available. Advantages of these arrays compared to the genome-wide SNP-arrays include direct measurement of high priority candidate DNA variants, also of low frequency, and lower costs. However, SNP density of these arrays is low and it is very likely that not all functional gene variation that contributes to the trait of interest is covered; evaluation of candidate gene DNA variants that are not on these SNP-arrays via direct measurement or tagging approaches is warranted. A more thorough selection of potentially interesting folate-related SNPs was performed by us and is described in chapter 5 of this thesis. An example of a comprehensive SNP selection procedure for a custom cardiovascular gene-centric array was described recently; the selection was based on, among others, results from candidate-gene and genome-wide association studies for CVD traits and intermediate phenotypes, pathway tools, and expression profiling studies, and resulted in a content-focused array containing almost 50,000 SNPs[164].

Utilization of large-scale genome sequence data

In the past few years, sequencing technologies have rapidly improved. The current next-generation sequencing technologies are capable to sequence DNA at high speed and reasonable cost and can facilitate the elucidation of the role of DNA variation in complex disease traits. These techniques have already been successfully applied on small scale to completely resequence individual genomes, resequence targeted genomic regions, and for the discovery of inherited and acquired structural variation[165]. Application of these new techniques to large-scale population studies especially offers new ways to investigate low-frequency and structural variants in positional or candidate regions. The data produced by the 1000 Genomes Project, in which the genomes of at least a thousand people from around the world will be sequenced, can be used to obtain high-resolution insight into genome variation that is present in ethnically diverse populations and may affect disease traits[166].

Incorporation of epigenetic and gene expression data

In addition to DNA variation, epigenetic variation, such as methylation or histone modification, may contribute to the heritability of one-carbon metabolism related traits. Hence, investigation of epigenetic effects is warranted when aiming for complete understanding of the molecular mechanisms underlying plasma tHcy and related traits and diseases.

Next to genome annotation information, knowledge about the functional consequences of DNA variations on gene expression and activity has been used in the past to select candidate polymorphisms and affirm the potential role of DNA variants in pathogenic mechanisms that link genes to phenotypic traits. However, information for a number of one-carbon metabolism candidate DNA variants is still lacking[167]. In addition, future positional candidates may be evaluated for their influence on gene expression as well. Also, identification of regulatory DNA variation on a regional or genome-wide scale may contribute to elucidation of disease pathology[168-170]. The development of infrastructure and statistical tools will be of importance to allow for efficient combination and utilization of these different layers of information (i.e. DNA variants, epigenetic variation, transcripts, metabolites, and disease).

Application of advanced genetic epidemiological designs and analyses

In congenital disorders like NTD in which the influence of maternal as well as offspring genotype effects are likely to play a role, the application of study designs that allow powerful differentiation between these effects is warranted to improve the characterization as well as the identification of genetic risk factors.

Valid en powerful evaluation of multi-locus effects can be facilitated by use of advanced techniques of analysis. The importance of the development, evaluation, and implementation of multi-locus techniques that can deal with large number of DNA

variants and environmental factors is recognized and ongoing[171]. The development of user-friendly software packages will facilitate real applications of these new techniques.

In addition, future genetic studies, in which multiple correlated one-carbon metabolism phenotypes are measured, may profit from analytical strategies that exploit the existing cross-trait covariance and reduce the multiple testing burden. For instance, multivariate tests of association[141] and principal components analyses[140] may be applied.

Also, incorporation of prior biological knowledge (e.g. enzyme kinetics, biological interaction of molecules) about one-carbon metabolism and related pathways in statistical genetic analysis as recently proposed by several researchers[172-174] may improve power and decrease false positive results.

Genetic epidemiological studies contribute to the understanding of complex disease aetiology. Identification and characterization of susceptibility genes facilitates development of diagnostic, prognostic, preventive and therapeutic tools. Plasma tHcy concentration has been viewed as important intermediate in clinical aetiology. It has especially raised much interest due to the fact that this one-carbon metabolism intermediate is modifiable by diet and B-vitamin supplementation. The potential decrease of vascular disease burden by homocysteine-lowering therapy is currently under debate, however. Identification of robust, strong genetic predictors of plasma tHcy will contribute to knowledge on whether plasma tHcy is indeed a causal intermediate or a bystander in disease aetiology and may shine light on the disappointing results of the homocysteine-lowering trials for vascular disease. Genetic epidemiological studies for a wider variety of one-carbon metabolism biochemical parameters are warranted and may identify new intermediates for disease and result in new insights into therapeutic and preventive measures. Evaluation of gene-environment interactions may reveal which groups respond to certain nutritional interventions and which not. To enable all this in coming years, the exploitation of new measurement techniques and the generation and utilization of vast amounts of genetic and phenotypic data is required, as well as the application of study designs and methods of analysis that can deal with the increasing complexity of the data at hand.

References

1. Gellekink H, den Heijer M, Heil SG, Blom HJ. Genetic determinants of plasma total homocysteine. Semin Vasc Med. 2005;5:98-109.
2. Ulvik A, Ueland PM, Fredriksen A, Meyer K, Vollset SE, Hoff G, Schneede J. Functional inference of the methylenetetrahydrofolate reductase 677C > T and 1298A > C polymorphisms from a large-scale epidemiological study. Hum Genet. 2007;121:57-64.

3. Fredriksen A, Meyer K, Ueland PM, Vollset SE, Grotmol T, Schneede J. Large-scale population-based metabolic phenotyping of thirteen genetic polymorphisms related to one-carbon metabolism. Hum Mutat. 2007;28:856-865.
4. Halsted CH, Wong DH, Peerson JM, Warden CH, Refsum H, Smith AD, Nygård OK, Ueland PM, Vollset SE, Tell GS. Relations of glutamate carboxypeptidase II (GCPII) polymorphisms to folate and homocysteine concentrations and to scores of cognition, anxiety, and depression in a homogeneous Norwegian population: the Hordaland Homocysteine Study. Am J Clin Nutr. 2007;86:514-521.
5. Collings A, Raitakari OT, Juonala M, Rontu R, Kähönen M, Hutri-Kähönen N, Rönnemaa T, Marniemi J, Viikari JS, Lehtimäki T. Associations of methylenetetrahydrofolate reductase C677T polymorphism with markers of subclinical atherosclerosis: the Cardiovascular Risk in Young Finns Study. Scand J Clin Lab Invest. 2008;68:22-30.
6. Golledge J, Norman PE. Relationship between two sequence variations in the gene for peroxisome proliferator-activated receptor-gamma and plasma homocysteine concentration. Health in men study. Hum Genet. 2008;123:35-40.
7. Yang QH, Botto LD, Gallagher M, Friedman JM, Sanders CL, Koontz D, Nikolova S, Erickson JD, Steinberg K. Prevalence and effects of gene-gene and gene-nutrient interactions on serum folate and serum total homocysteine concentrations in the United States: findings from the third National Health and Nutrition Examination Survey DNA Bank. Am J Clin Nutr. 2008;88:232-246.
8. Zee RY, Mora S, Cheng S, Erlich HA, Lindpaintner K, Rifai N, Buring JE, Ridker PM. Homocysteine, 5,10-methylenetetrahydrofolate reductase 677C>T polymorphism, nutrient intake, and incident cardiovascular disease in 24,968 initially healthy women. Clin Chem. 2007;53:845-851.
9. Lwin H, Yoshiike N, Yokoyama T, Saito K, Date C, Tanaka H. The relationships between plasma total homocysteine and selected atherosclerotic risk factors according to the C677T methylenetetrahydrofolate reductase gene in Japanese. Eur J Cardiovasc Prev Rehabil. 2005;12:182-184.
10. Muntjewerff JW, Hoogendoorn ML, Kahn RS, Sinke RJ, Den Heijer M, Kluijtmans LA, Blom HJ. Hyperhomocysteinemia, methylenetetrahydrofolate reductase 677TT genotype, and the risk for schizophrenia: a Dutch population based case-control study. Am J Med Genet B Neuropsychiatr Genet. 2005;135B:69-72.
11. Guéant-Rodriguez RM, Juilliére Y, Candito M, Adjalla CE, Gibelin P, Herbeth B, Van Obberghen E, Gueánt JL. Association of MTRRA66G polymorphism (but not of MTHFR C677T and A1298C, MTRA2756G, TCN C776G) with homocysteine and coronary artery disease in the French population. Thromb Haemost. 2005;94:510-515.
12. Sofi F, Marcucci R, Giusti B, Pratesi G, Lari B, Sestini I, Lo Sapio P, Pulli R, Pratesi C, Abbate R, Gensini GF. High levels of homocysteine, lipoprotein (a) and plasminogen activator inhibitor-1 are present in patients with abdominal aortic aneurysm. Thromb Haemost. 2005;94:1094-1098.

13. Kebert CB, Eichner JE, Moore WE, Schechter E, Yaoi T, Vogel S, Allen RA, Dunn ST. Relationship of the 1793G-A and 677C-T polymorphisms of the 5,10-methylenetetrahydrofolate reductase gene to coronary artery disease. Dis Markers. 2006;22:293-301.
14. Ho GY, Eikelboom JW, Hankey GJ, Wong CR, Tan SL, Chan JB, Chen CP. Methylenetetrahydrofolate reductase polymorphisms and homocysteine-lowering effect of vitamin therapy in Singaporean stroke patients. Stroke. 2006;37:456-60.
15. Pereira AC, Miyakawa AA, Lopes NH, Soares PR, de Oliveira SA, Cesar LA, Ramires JF, Hueb W, Krieger JE. Dynamic regulation of MTHFR mRNA expression and C677T genotype modulate mortality in coronary artery disease patients after revascularization. Thromb Res. 2007;121:25-32.
16. Mtiraoui N, Ezzidi I, Chaieb M, Marmouche H, Aouni Z, Chaieb A, Mahjoub T, Vaxillaire M, Almawi WY. MTHFR C677T and A1298C gene polymorphisms and hyperhomocysteinemia as risk factors of diabetic nephropathy in type 2 diabetes patients. Diabetes Res Clin Pract. 2007;75:99-106.
17. Wang L, Ke Q, Chen W, Wang J, Tan Y, Zhou Y, Hua Z, Ding W, Niu J, Shen J, Zhang Z, Wang X, Xu Y, Shen H. Polymorphisms of MTHFD, plasma homocysteine levels, and risk of gastric cancer in a high-risk Chinese population. Clin Cancer Res. 2007;13:2526-2532.
18. Siva A, De Lange M, Clayton D, Monteith S, Spector T, Brown MJ. The heritability of plasma homocysteine, and the influence of genetic variation in the homocysteine methylation pathway. QJM. 2007;100:495-499.
19. Liu CS, Chen CH, Chiang HC, Kuo CL, Huang CS, Cheng WL, Wei YH, Chen HW. B-group vitamins, MTHFR C677T polymorphism and carotid intima-media thickness in clinically healthy subjects. Eur J Clin Nutr. 2007;61:996-1003.
20. Vollset SE, Igland J, Jenab M, Fredriksen A, Meyer K, Eussen S, Gjessing HK, Ueland PM, Pera G, Sala N, Agudo A, Capella G, Del Giudice G, Palli D, Boeing H, Weikert C, Bueno-de-Mesquita HB, Carneiro F, Pala V, Vineis P, Tumino R, Panico S, Berglund G, Manjer J, Stenling R, Hallmans G, Martínez C, Dorronsoro M, Barricarte A, Navarro C, Quirós JR, Allen N, Key TJ, Bingham S, Linseisen J, Kaaks R, Overvad K, Tjønneland A, Büchner FL, Peeters PH, Numans ME, Clavel-Chapelon F, Boutron-Ruault MC, Trichopoulou A, Lund E, Slimani N, Ferrari P, Riboli E, González CA. The association of gastric cancer risk with plasma folate, cobalamin, and methylenetetrahydrofolate reductase polymorphisms in the European Prospective Investigation into Cancer and Nutrition. Cancer Epidemiol Biomarkers Prev. 2007;16:2416-2424.
21. Yazdanpanah N, Uitterlinden AG, Zillikens MC, Jhamai M, Rivadeneira F, Hofman A, de Jonge R, Lindemans J, Pols HA, van Meurs JB. Low dietary riboflavin but not folate predicts increased fracture risk in postmenopausal women homozygous for the MTHFR 677 T allele. J Bone Miner Res. 2008;23:86-94.
22. Ghazouani L, Abboud N, Mtiraoui N, Zammiti W, Addad F, Amin H, Almawi WY, Mahjoub T. Homocysteine and methylenetetrahydrofolate reductase C677T and A1298C polymorphisms in Tunisian patients with severe coronary artery disease. J Thromb Thrombolysis. 2009;27:191-197.

23. Freitas AI, Mendonça I, Guerra G, Brión M, Reis RP, Carracedo A, Brehm A. Methylenetetrahydrofolate reductase gene, homocysteine and coronary artery disease: The A1298C polymorphism does matter. Inferences from a case study (Madeira, Portugal). Thromb Res. 2008;122:648-656.
24. Kim JM, Stewart R, Kim SW, Yang SJ, Shin IS, Shin HY, Yoon JS. Methylenetetrahydrofolate reductase gene and risk of Alzheimer's disease in Koreans. Int J Geriatr Psychiatry. 2008;23:454-459.
25. Ozbek Z, Kucukali CI, Ozkok E, Orhan N, Aydin M, Kilic G, Sazci A, Kara I. Effect of the methylenetetrahydrofolate reductase gene polymorphisms on homocysteine, folate and vitamin B12 in patients with bipolar disorder and relatives. Prog Neuropsychopharmacol Biol Psychiatry. 2008;32:1331-1337.
26. Fatini C, Sofi F, Gori AM, Sticchi E, Marcucci R, Lenti M, Casini A, Surrenti C, Abbate R, Gensini GF. Endothelial nitric oxide synthase -786T>C, but not 894G>T and 4a4b, polymorphism influences plasma homocysteine concentrations in persons with normal vitamin status. Clin Chem. 2005;51:1159-1164.
27. Chiuve SE, Giovannucci EL, Hankinson SE, Hunter DJ, Stampfer MJ, Willett WC, Rimm EB. Alcohol intake and methylenetetrahydrofolate reductase polymorphism modify the relation of folate intake to plasma homocysteine. Am J Clin Nutr. 2005;82:155-162.
28. Lim U, Peng K, Shane B, Stover PJ, Litonjua AA, Weiss ST, Gaziano JM, Strawderman RL, Raiszadeh F, Selhub J, Tucker KL, Cassano PA. Polymorphisms in cytoplasmic serine hydroxymethyltransferase and methylenetetrahydrofolate reductase affect the risk of cardiovascular disease in men. J Nutr. 2005;135:1989-1994.
29. Guéant-Rodriguez RM, Guéant JL, Debard R, Thirion S, Hong LX, Bronowicki JP, Namour F, Chabi NW, Sanni A, Anello G, Bosco P, Romano C, Amouzou E, Arrieta HR, Sánchez BE, Romano A, Herbeth B, Guilland JC, Mutchinick OM. Prevalence of methylenetetrahydrofolate reductase 677T and 1298C alleles and folate status: a comparative study in Mexican, West African, and European populations. Am J Clin Nutr. 2006;83:701-707.
30. Devlin AM, Clarke R, Birks J, Evans JG, Halsted CH. Interactions among polymorphisms in folate-metabolizing genes and serum total homocysteine concentrations in a healthy elderly population. Am J Clin Nutr. 2006;83:708-713.
31. Martínez ME, Thompson P, Jacobs ET, Giovannucci E, Jiang R, Klimecki W, Alberts DS. Dietary factors and biomarkers involved in the methylenetetrahydrofolate reductase genotype-colorectal adenoma pathway. Gastroenterology. 2006;131:1706-1716.
32. Giusti B, Gori AM, Marcucci R, Sestini I, Saracini C, Sticchi E, Gensini F, Fatini C, Abbate R, Gensini GF. Role of C677T and A1298C MTHFR, A2756G MTR and -786 C/T eNOS gene polymorphisms in atrial fibrillation susceptibility. PLoS ONE. 2007;2:e495.
33. Ndrepepa G, Kastrati A, Braun S, Koch W, Kölling K, Mehilli J, Schömig A. Circulating homocysteine levels in patients with type 2 diabetes mellitus. Nutr Metab Cardiovasc Dis. 2008;18:66-73.

34. Naess IA, Christiansen SC, Romundstad PR, Cannegieter SC, Blom HJ, Rosendaal FR, Hammerstrøm J. Prospective study of homocysteine and MTHFR 677TT genotype and risk for venous thrombosis in a general population--results from the HUNT 2 study. Br J Haematol. 2008;141:529-535.
35. Bathum L, Petersen I, Christiansen L, Konieczna A, Sørensen TI, Kyvik KO. Genetic and environmental influences on plasma homocysteine: results from a Danish twin study. Clin Chem. 2007;53:971-979.
36. Eklof V, Van Guelpen B, Hultdin J, Johansson I, Hallmans G, Palmqvist R. The reduced folate carrier (RFC1) 80G>A and folate hydrolase 1 (FOLH1) 1561C>T polymorphisms and the risk of colorectal cancer: a nested case-referent study. Scand J Clin Lab Invest. 2007;21:1-9.
37. Dufficy L, Naumovski N, Ng X, Blades B, Yates Z, Travers C, Lewis P, Sturm J, Veysey M, Roach PD, Lucock MD. G80A reduced folate carrier SNP influences the absorption and cellular translocation of dietary folate and its association with blood pressure in an elderly population. Life Sci. 2006;79:957-966.
38. Gellekink H, Blom HJ, den Heijer M. Associations of common polymorphisms in the thymidylate synthase, reduced folate carrier and 5-aminoimidazole-4-carboxamide ribonucleotide transformylase/inosine monophosphate cyclohydrolase genes with folate and homocysteine levels and venous thrombosis risk. Clin Chem Lab Med. 2007;45:471-476.
39. Biselli JM, Goloni-Bertollo EM, Haddad R, Eberlin MN, Pavarino-Bertelli EC. The MTR A2756G polymorphism is associated with an increase of plasma homocysteine concentration in Brazilian individuals with Down syndrome. Braz J Med Biol Res. 2008;41:34-40.
40. Barbosa PR, Stabler SP, Trentin R, Carvalho FR, Luchessi AD, Hirata RD, Hirata MH, Allen RH, Guerra-Shinohara EM. Evaluation of nutritional and genetic determinants of total homocysteine, methylmalonic acid and S-adenosylmethionine/S-adenosylhomocysteine values in Brazilian childbearing-age women. Clin Chim Acta. 2008;388:139-147.
41. Guéant JL, Anello G, Bosco P, Guéant-Rodríguez RM, Romano A, Barone C, Gérard P, Romano C. Homocysteine and related genetic polymorphisms in Down's syndrome IQ. J Neurol Neurosurg Psychiatry. 2005;76:706-709.
42. Aléssio AC, Höehr NF, Siqueira LH, Bydlowski SP, Annichino-Bizzacchi JM. Polymorphism C776G in the transcobalamin II gene and homocysteine, folate and vitamin B12 concentrations. Association with MTHFR C677T and A1298C and MTRR A66G polymorphisms in healthy children.Thromb Res. 2007;119:571-577.
43. Guéant JL, Chabi NW, Guéant-Rodriguez RM, Mutchinick OM, Debard R, Payet C, Lu X, Villaume C, Bronowicki JP, Quadros EV, Sanni A, Amouzou E, Xia B, Chen M, Anello G, Bosco P, Romano C, Arrieta HR, Sánchez BE, Romano A, Herbeth B, Anwar W, Namour F. Environmental influence on the worldwide prevalence of a 776C->G variant in the transcobalamin gene (TCN2). J Med Genet. 2007;44:363-367.
44. Brouns R, Ursem N, Lindemans J, Hop W, Pluijm S, Steegers E, Steegers-Theunissen R. Polymorphisms in genes related to folate and cobalamin metabolism and the associations with complex birth defects. Prenat Diagn. 2008;28:485-493.

45. Lievers KJ, Kluijtmans LA, Heil SG, Boers GH, Verhoef P, van Oppenraay-Emmerzaal D, den Heijer M. Trijbels FJ, Blom HJ. A 31 bp VNTR in the cystathionine beta-synthase (CBS) gene is associated with reduced CBS activity and elevated post-load homocysteine levels. Eur J Hum Genet 2001;9:583-589.
46. Lievers KJ, Kluijtmans LA, Blom HJ, Wilson PW, Selhub J, Ordovas JM. Association of a 31 bp VNTR in the CBS gene with postload homocysteine concentrations in the Framingham Offspring Study. Eur J Hum Genet. 2006;14:1125-1129.
47. Goodman JE, Lavigne JA, Wu K, Helzlsouer KJ, Strickland PT, Selhub J, Yager JD. COMT genotype, micronutrients in the folate metabolic pathway and breast cancer risk. Carcinogenesis. 2001;22:1661-1665.
48. Geisel J, Hübner U, Bodis M, Schorr H, Knapp JP, Obeid R, Herrmann W. The role of genetic factors in the development of hyperhomocysteinemia. Clin Chem Lab Med. 2003;41:1427-1434.
49. Gellekink H, Muntjewerff JW, Vermeulen SH, Hermus AR, Blom HJ, den Heijer M. Catechol-O-methyltransferase genotype is associated with plasma total homocysteine levels and may increase venous thrombosis risk. Thromb Haemost. 2007;98:1226-1231.
50. Voutilainen S, Tuomainen TP, Korhonen M, Mursu J, Virtanen JK, Happonen P, Alfthan G, Erlund I, North KE, Mosher MJ, Kauhanen J, Tiihonen J, Kaplan GA, Salonen JT. Functional COMT Val158Met polymorphism, risk of acute coronary events and serum homocysteine: the kuopio ischaemic heart disease risk factor study. PLoS ONE. 2007;2:e181.
51. Tunbridge EM, Harrison PJ, Warden DR, Johnston C, Refsum H, Smith AD. Polymorphisms in the catechol-O-methyltransferase (COMT) gene influence plasma total homocysteine levels. Am J Med Genet B Neuropsychiatr Genet. 2008;147B:996-999.
52. Zammiti W, Mtiraoui N, Mahjoub T. Lack of consistent association between endothelial nitric oxide synthase gene polymorphisms, homocysteine levels and recurrent pregnancy loss in tunisian women. Am J Reprod Immunol. 2008;59:139-145.
53. Souto JC, Blanco-Vaca F, Soria JM, Buil A, Almasy L, Ordoñez-Llanos J, Martín-Campos JM, Lathrop M, Stone W, Blangero J, Fontcuberta J. A genomewide exploration suggests a new candidate gene at chromosome 11q23 as the major determinant of plasma homocysteine levels: results from the GAIT project. Am J Hum Genet. 2005;76:925-933.
54. Zhang L, Miyaki K, Araki J, Nakayama T, Muramatsu M. The relation between nicotinamide N-methyltransferase gene polymorphism and plasma homocysteine concentration in healthy Japanese men. Thromb Res. 2007;121:55-58.
55. Shin BS, Oh SY, Kim YS, Kim KW. The paraoxonase gene polymorphism in stroke patients and lipid profile. Acta Neurol Scand. 2008;117:237-243.
56. Gellekink H, Blom HJ, van der Linden IJ, den Heijer M. Molecular genetic analysis of the human dihydrofolate reductase gene: relation with plasma total homocysteine, serum and red blood cell folate levels. Eur J Hum Genet. 2007;15:103-109.

57. Stanisławska-Sachadyn A, Brown KS, Mitchell LE, Woodside JV, Young IS, Scott JM, Murray L, Boreham CA, McNulty H, Strain JJ, Whitehead AS. An insertion/deletion polymorphism of the dihydrofolate reductase (DHFR) gene is associated with serum and red blood cell folate concentrations in women. Hum Genet. 2008;123:289-295.
58. Vermeulen SH, van der Vleuten GM, de Graaf J, Hermus AR, Blom HJ, Stalenhoef AF, den Heijer M. A genome-wide linkage scan for homocysteine levels suggests three regions of interest. J Thromb Haemost. 2006;4:1303-1307.
59. Kullo IJ, Ding K, Boerwinkle E, Turner ST, Mosley TH Jr, Kardia SL, de Andrade M. Novel genomic loci influencing plasma homocysteine levels. Stroke. 2006;37:1703-1709.
60. den Heijer M, Graafsma S, Lee SY, van Landeghem B, Kluijtmans L, Verhoef P, Beaty TH, Blom H. Homocysteine levels--before and after methionine loading--in 51 Dutch families. Eur J Hum Genet. 2005;13:753-762.
61. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. Lancet. 1995;346:1070-1071.
62. Lievers KJ, Boers GH, Verhoef P, den Heijer M, Kluijtmans LA, van der Put NM, Trijbels FJ, Blom HJ. A second common variant in the methylenetetrahydrofolate reductase (MTHFR) gene and its relationship to MTHFR enzyme activity, homocysteine, and cardiovascular disease risk. J Mol Med. 2001;79:522-528.
63. Kluijtmans LA, Boers GH, Trijbels FJ, van Lith-Zanders HM, van den Heuvel LP, Blom HJ. A common 844INS68 insertion variant in the cystathionine beta-synthase gene. Biochem Mol Med. 1997;62:23-25.
64. Ma J, Stampfer MJ, Hennekens CH, Frosst P, Selhub J, Horsford J, Malinow MR, Willett WC, Rozen R. Methylenetetrahydrofolate reductase polymorphism, plasma folate, homocysteine, and risk of myocardial infarction in US physicians. Circulation. 1996;94:2410-2416.
65. Davey Smith G, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? BMJ. 2005;330:1076-1079.
66. Kluijtmans LA, Young IS, Boreham CA, Murray L, McMaster D, McNulty H, Strain JJ, McPartlin J, Scott JM, Whitehead AS. Genetic and nutritional factors contributing to hyperhomocysteinemia in young adults. Blood. 2003;101:2483-2488.
67. Lucock M, Yates Z. Synergy between 677 TT MTHFR genotype and related folate SNPs regulates homocysteine level. Nutrition Research. 2006;26:180-185.
68. Afman LA, Lievers KJ, Kluijtmans LA, Trijbels FJ, Blom HJ. Gene-gene interaction between the cystathionine beta-synthase 31 base pair variable number of tandem repeats and the methylenetetrahydrofolate reductase 677C > T polymorphism on homocysteine levels and risk for neural tube defects. Mol Genet Metab. 2003;78:211-215.
69. Ebrahim S, Davey Smith G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? Hum Genet. 2008;123:15-33.
70. van der Linden I, Afman LA, Heil SG, Blom HJ. Genetic variation in genes of folate metabolism and neural-tube defect risk. Proc Nutr Soc. 2006;65:204-215.

71. Ray JG, Blom HJ. Vitamin B12 insufficiency and the risk of fetal neural tube defects. QJM 2003;96:289-295.
72. Groenen PM, van Rooij, I, Peer PG, Gooskens RH, Zielhuis GA, Steegers-Theunissen RP. Marginal maternal vitamin B12 status increases the risk of offspring with spina bifida. Am J Obstet Gynecol 2004;191:11-17.
73. Blom HJ, Shaw GM, den Heijer M, Finnell RH. Neural tube defects and folate: case far from closed. Nat Rev Neurosci. 2006;7:724-731.
74. den Heijer M, Lewington S, Clarke R. (2005) Homocysteine, MTHFR and risk of venous thrombosis: a meta-analysis of published epidemiological studies. J Thromb Haemost 2005;3:292-299
75. Lentz SR. Mechanisms of homocysteine-induced atherothrombosis. J Thromb Haemost. 2005;3:1646-1654.
76. Nedergaard J, Cannon B.The 'novel' 'uncoupling' proteins UCP2 and UCP3: what do they really do? Pros and cons for suggested functions. Exp Physiol. 2003;88:65-84.
77. Blanc J, Alves-Guerra MC, Esposito B, Rousset S, Gourdy P, Ricquier D, Tedgui A, Miroux B, Mallat Z. Protective role of uncoupling protein 2 in atherosclerosis. Circulation. 2003;107:388-390.
78. Dhamrait SS, Stephens JW, Cooper JA, Acharya J, Mani AR, Moore K, Miller GJ, Humphries SE, Hurel SJ, Montgomery HE. Cardiovascular risk in healthy men and markers of oxidative stress in diabetic men are associated with common variation in the gene for uncoupling protein 2. Eur Heart J. 2004;25:468-475.
79. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. Eur J Hum Genet. 2004;12:964-970.
80. Putter H, Houwing-Duistermaat JJ, Nagelkerke NJD. Combining evidence for association from transmission disequilibrium and case-control studies using single-nucleotide polymorphisms. BMC Genetics. 2005;6:S106.
81. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet. 2004;76:592-608.
82. Starr JR, Hsu L, Schwartz SM. Assessing maternal genetic associations: a comparison of the log-linear approach to case-parent triad data and a case-control approach. Epidemiology. 2005;16:294-303.
83. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet. 1998;62:969-978.
84. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet. 2005;77:627-636.
85. Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. Am J Epidemiol. 2008;168:541-547.

86. Ickstadt K, Schäfer M, Fritsch A, Schwender H, Abel J, Bolt HM, Brüning T, Ko YD, Vetter H, Harth V. Statistical methods for detecting genetic interactions: a head and neck squamous-cell cancer study. Toxicol Environ Health A. 2008;71:803-815.
87. Andrew AS, Karagas MR, Nelson HH, Guarrera S, Polidoro S, Gamberini S, Sacerdote C, Moore JH, Kelsey KT, Demidenko E, Vineis P, Matullo G. DNA repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy. Hum Hered. 2008;65:105-118.
88. Justenhoven C, Hamann U, Schubert F, Zapatka M, Pierl CB, Rabstein S, Selinski S, Mueller T, Ickstadt K, Gilbert M, Ko YD, Baisch C, Pesch B, Harth V, Bolt HM, Vollmert C, Illig T, Eils R, Dippon J, Brauch H. Breast cancer: a candidate gene approach across the estrogen metabolic pathway. Breast Cancer Res Treat. 2008;108:137-149.
89. Infante J, García-Gorostiaga I, Sánchez-Juan P, Sánchez-Quintana C, Gurpegui JL, Rodríguez-Rodríguez E, Mateo I, Berciano J, Combarros O. Inflammation-related genes and the risk of Parkinson's disease: a multilocus approach. Eur J Neurol. 2008;15:431-433.
90. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001;69:138-147.
91. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003;24:150-157.
92. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet Epidemiol. 2001;21 Suppl 1:S626-S631.
93. Ruczinski I, Kooperberg C, LeBlanc L. Exploring interactions in high-dimensional genomic data: an overview of Logic Regression, with applications. J Multiv Anal. 2004;90:178-195.
94. Hoh J, Wille A, Ott J. Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res. 2001;11:2115-2119.
95. Wille A, Hoh J, Ott J. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol. 2003;25:350-359.
96. Mei H, Cuccaro ML, Martin ER. Multifactor dimensionality reduction-phenomics: a novel method to capture genetic heterogeneity with use of phenotypic variables. Am J Hum Genet. 2007;81:1251-1261.
97. Heidema AG, Feskens EJ, Doevendans PA, Ruven HJ, van Houwelingen HC, Mariman EC, Boer JM. Analysis of multiple SNPs in genetic association studies: comparison of three multi-locus methods to prioritize and select SNPs. Genet Epidemiol. 2007;31:910-921.
98. Briollais L, Wang Y, Rajendram I, Onay V, Shi E, Knight J, Ozcelik H. Methodological issues in detecting gene-gene interactions in breast cancer susceptibility: a population-based study in Ontario. BMC Med. 2007;5:22.
99. Bush WS, Dudek SM, Ritchie MD. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. Bioinformatics. 2006;22:2173-2174.
100. Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol. 2007;31:306-315.

101. Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics. 2008;9:238.
102. Lou XY, Chen GB, Yan L, Ma JZ, Zhu J, Elston RC, Li MD. A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. Am J Hum Genet. 2007;80:1125-1137.
103. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol. 2005;28:157-170.
104. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. Biostatistics. 2008;9:187-198.
105. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet. 2006;79:1002-1016.
106. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516-1517.
107. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-945.
108. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299-1320.
109. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851-861.
110. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest. 2008;118:1590-1605.
111. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Pálsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Vermeulen SH, Goldstein AM, Tucker MA, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet. 2008;40:835-837.
112. Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, Vermeulen SH, Hulsbergen-van de Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, Rudnai P, Kellen E, Polidoro S, Guarrera S, Sacerdote C, Sanchez M, Saez B, Valdivia G, Ryk C, de Verdier P, Lindblom A, Golka K, Bishop DT, Knowles MA, Nikulasson S, Petursdottir V, Jonsson E, Geirsson G, Kristjansson B, Mayordomo JI, Steineck G, Porru S, Buntinx F, Zeegers MP, Fletcher T, Kumar R, Matullo G, Vineis P, Kiltie AE, Gulcher JR, Thorsteinsdottir U, Kong A, Rafnar T, Stefansson K. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet. 2008;40:1307-1312.

113. Hattersley AT, McCarthy MI. What makes a good genetic association study? Lancet. 2005;366:1315-1323.
114. Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet. 2002;71:304-311.
115. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev. 2002;11:513-520.
116. Khlat M, Cazes MH, Génin E, Guiguet M. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. Cancer Epidemiol Biomarkers Prev. 2004;13:1660-1664.
117. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol. 2001;60:227-237.
118. Harris MJ, Juriloff DM. Mini-review: toward understanding mechanisms of genetic neural tube defects in mice. Teratology. 1999:60:292-305.
119. Relton CL, Wilding CS, Jonas PA, Lynch SA, Tawn EJ, Burn J. Genetic susceptibility to neural tube defect pregnancy varies with offspring phenotype. Clin Genet. 2003;64:424-428.
120. Park CH, Stewart W, Khoury MJ, Mulinare J. Is there etiologic heterogeneity between upper and lower neural tube defects? Am J Epidemiol. 1992;136:1493-1501.
121. van Stralen KJ, Doggen CJ, Bezemer ID, Pomp ER, Lisman T, Rosendaal FR. Mechanisms of the factor V Leiden paradox. Arterioscler Thromb Vasc Biol. 2008;28:1872-1877.
122. te Poele-Pothoff MT, van den Berg M, Franken DG, Boers GH, Jakobs C, de Kroon IF, Eskes TK, Trijbels JM, Blom HJ. Three different methods for the determination of total homocysteine in plasma. Ann Clin Biochem. 1995;32:218-220.
123. Refsum H, Smith AD, Ueland PM, Nexo E, Clarke R, McPartlin J, Johnston C, Engbaek F, Schneede J, McPartlin C, Scott JM. Facts and recommendations about total homocysteine determinations: an expert opinion. Clin Chem. 2004;50:3-32.
124. Boushey CJ, Beresford SA, Omenn GS, Motulsky AG. A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. Probable benefits of increasing folic acid intakes. JAMA. 1995;274:1049-1057.
125. Graham IM, Daly LE, Refsum HM, Robinson K, Brattström LE, Ueland PM, Palma-Reis RJ, Boers GH, Sheahan RG, Israelsson B, Uiterwaal CS, Meleady R, McMaster D, Verhoef P, Witteman J, Rubba P, Bellet H, Wautrecht JC, de Valk HW, Sales Lúis AC, Parrot-Rouland FM, Tan KS, Higgins I, Garcon D, Andria G, et al. Plasma homocysteine as a risk factor for vascular disease. The European Concerted Action Project. JAMA. 1997;277:1775-1781.
126. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global variation in copy number in the human genome. Nature. 2006;444:444-454.

127. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007;39:S37-42.
128. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007;315:848-853.
129. Human Genome Structural Variation Working Group, Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, Carter NP, Church DM, Felsenfeld A, Guyer M, Lee C, Lupski JR, Mullikin JC, Pritchard JK, Sebat J, Sherry ST, Smith D, Valle D, Waterston RH. Completing the map of human genetic variation. Nature. 2007;447:161-165.
130. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Genet Rev. 2002;3:1-7.
131. Zou GY, Donner A.The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann Hum Genet. 2006;70:923-933.
132. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet. 2004;12:395-399.
133. Purcell S, Daly MJ, Sham PC. WHAP: haplotype-based association analysis. Bioinformatics. 2007;23:255-256.
134. Schaid DJ. Evaluating associations of haplotypes with traits. Genet Epidemiol. 2004;27:348-364.
135. Cordell HJ. Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. Genet Epidemiol. 2006;30:259-275.
136. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087-1091.
137. Yu Z, Schaid DJ. Methods to impute missing genotypes for population data. Hum Genet. 2007;122:495-504.
138. Almasy L, Dyer TD, Blangero J. Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genet Epidemiol. 1997;14:953-958.
139. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B. Murphy A, Silverman EK, MacGregor A, Weiss ST, Laird NM. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Stat Appl Genet Mol Biol. 2004;3:Article17.
140. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet Epidemiol. 2008;32:9-19.
141. Ferreira MA, Purcell SM. A multivariate test of association. Bioinformatics. 2009;25:132-133.

142. Bazzano LA, Reynolds K, Holder KN, He J. Effect of folic acid supplementation on risk of cardiovascular diseases: a meta-analysis of randomized controlled trials. JAMA. 2006;296:2720-2726.
143. Ebbing M, Bleie Ø, Ueland PM, Nordrehaug JE, Nilsen DW, Vollset SE, Refsum H, Pedersen EK, Nygård O. Mortality and cardiovascular events in patients treated with homocysteine-lowering B vitamins after coronary angiography: a randomized controlled trial. JAMA. 2008;300:795-804.
144. Wang X, Qin X, Demirtas H, Li J, Mao G, Huo Y, Sun N, Liu L, Xu X. Efficacy of folic acid supplementation in stroke prevention: a meta-analysis. Lancet. 2007;369:1876-1882.
145. den Heijer M, Willems HP, Blom HJ, Gerrits WB, Cattaneo M, Eichinger S, Rosendaal FR, Bos GM. Homocysteine lowering by B vitamins and the secondary prevention of deep vein thrombosis and pulmonary embolism: A randomized, placebo-controlled, double-blind trial. Blood. 2007;109:139-144.
146. Ray JG, Kearon C, Yi Q, Sheridan P, Lonn E; Heart Outcomes Prevention Evaluation 2 (HOPE-2) Investigators. Homocysteine-lowering therapy and risk for venous thromboembolism: a randomized trial. Ann Intern Med. 2007;146:761-767.
147. de Bree A, van Mierlo LA, Draijer R. Folic acid improves vascular reactivity in humans: a meta-analysis of randomized controlled trials. Am J Clin Nutr. 2007;86:610-617.
148. Potter K, Hankey GJ, Green DJ, Eikelboom J, Jamrozik K, Arnolda LF. The effect of long-term homocysteine-lowering on carotid intima-media thickness and flow-mediated vasodilation in stroke patients: a randomized controlled trial and meta-analysis. BMC Cardiovasc Disord. 2008;8:24.
149. Wald DS, Wald NJ, Morris JK, Law M. Folic acid, homocysteine, and cardiovascular disease: judging causality in the face of inconclusive trial evidence. BMJ. 2006;333:1114-1147.
150. B-Vitamin Treatment Trialists' Collaboration. Homocysteine-lowering trials for prevention of cardiovascular events: a review of the design and power of the large randomized trials. Am Heart J. 2006;151:282-287.
151. Ueland PM, Clarke R. Homocysteine and cardiovascular risk: considering the evidence in the context of study design, folate fortification, and statistical power. Clin Chem. 2007;53:807-809.
152. Gellekink H, van Oppenraaij-Emmerzaal D, van Rooij A, Struys EA, den Heijer M, Blom HJ. Stable-isotope dilution liquid chromatography-electrospray injection tandem mass spectrometry method for fast, selective measurement of S-adenosylmethionine and S-adenosylhomocysteine in plasma. Clin Chem. 2005;51:1487-1492.
153. Ueland PM, Midttun O, Windelberg A, Svardal A, Skålevik R, Hustad S. Quantitative profiling of folate and one-carbon metabolism in large-scale epidemiological studies by mass spectrometry. Clin Chem Lab Med. 2007;45:1737-1745
154. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009;38:263-273.

155. Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. Hum Hered. 2007;64:203-213.
156. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661-678.
157. Chinthammitr Y, Vos HL, Rosendaal FR, Doggen CJ. The association of prothrombin A19911G polymorphism with plasma prothrombin activity and venous thrombosis: results of the MEGA study, a large population-based case-control study. J Thromb Haemost. 2006;4:2587-2592.
158. Seminara D, Khoury MJ, O'Brien TR, Manolio T, Gwinn ML, Little J, Higgins JP, Bernstein JL, Boffetta P, Bondy M, Bray MS, Brenchley PE, Buffler PA, Casas JP, Chokkalingam AP, Danesh J, Davey Smith G, Dolan S, Duncan R, Gruis NA, Hashibe M, Hunter D, Jarvelin MR, Malmer B, Maraganore DM, Newton-Bishop JA, Riboli E, Salanti G, Taioli E, Timpson N, Uitterlinden AG, Vineis P, Wareham N, Winn DM, Zimmern R, Ioannidis JP. Human Genome Epidemiology Network; the Network of Investigator Networks. The emergence of networks in human genome epidemiology: challenges and opportunities. Epidemiology. 2007;18:1-8.
159. van Meurs J, Rivadeneira F, de Jonge R, Arp P, Jhamai M, Hofman A, Plos H, Lindemans J, Uitterlinden AG. Clin Chem and Lab Med. 2007;45:A19.
160. Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, Meitinger T, Mewes HW, Wichmann HE, Weinberger KM, Adamski J, Illig T, Suhre K. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 2008;4:e1000282.
161. Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, Chanock SJ, Hunter DJ. Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat Genet. 2008;40:1160-1162.
162. Meyer K, Fredriksen A, Ueland PM. High-level multiplex genotyping of polymorphisms involved in folate or homocysteine metabolism by matrix-assisted laser desorption/ionization mass spectrometry. Clin Chem. 2004;50:391-402.
163. Giusti B, Sestini I, Saracini C, Sticchi E, Bolli P, Magi A, Gori AM, Marcucci R, Gensini GF, Abbate R. High-throughput multiplex single-nucleotide polymorphism (SNP) analysis in genes involved in methionine metabolism. Biochem Genet. 2008;46:406-423.
164. Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, Galver L, Barrett JC, Grant SF, Farlow DN, Chandrupatla HR, Hansen M, Ajmal S, Papanicolaou GJ, Guo Y, Li M, Derohannessian S, de Bakker PI, Bailey SD, Montpetit A, Edmondson AC, Taylor K, Gai X, Wang SS, Fornage M, Shaikh T, Groop L, Boehnke M, Hall AS, Hattersley AT, Frackelton E, Patterson N, Chiang CW, Kim CE, Fabsitz RR, Ouwehand W, Price AL, Munroe P, Caulfield M, Drake T, Boerwinkle E, Reich D, Whitehead AS, Cappola TP, Samani NJ, Lusis AJ, Schadt E, Wilson JG, Koenig W, McCarthy MI, Kathiresan S, Gabriel SB, Hakonarson H, Anand SS, Reilly M, Engert JC, Nickerson DA, Rader DJ, Hirschhorn JN, Fitzgerald GA. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS ONE. 2008;3:e3583.

165. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135-1145.
166. Kuehn BM. 1000 Genomes Project promises closer look at variation in human genome. JAMA. 2008;300:2715.
167. Sharma P, Senthilkumar RD, Brahmachari V, Sundaramoorthy E, Mahajan A, Sharma A, Sengupta S. Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. Lipids Health Dis. 2006;5:1.
168. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17:R156-165.
169. Nica AC, Dermitzakis ET. Using gene expression to investigate the genetic basis of complex disorders. Hum Mol Genet. 2008;17:R129-134.
170. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. Genetics of gene expression and its effect on disease. Nature. 2008;452:423-428.
171. Manolio TA, Collins FS. Genes, Environment, Health, and Disease: Facing up to Complexity. Hum Hered. 2007;63:63-66.
172. Conti DV, Figueiredo JC, Levine AJ, Ulrich CM, Poynter JN, Nijhout HF, Reed M, Zheng Y, Haile RW. Using in silico priors for pathway modeling. Genet Epidemiol. 2008;32:669.
173. Volk HE, Gilliland F, Diaz-Sanchez D, Conti DV. Hierarchical Modeling of Pathway-Based Candidate Genes and Gene-Environment Interactions Genet Epidemiol. 2008;32:669.
174. Bush WS, Dudek SM, Haines JL, Ritchie MD. Candidate Epistasis: Generating putative gene-gene interactions for an analysis of a whole-genome association study of multiple sclerosis. Genet Epidemiol. 2008;32:670.

Summary

Homocysteine is a sulphur-containing intermediate in one-carbon metabolism. Plasma total homocysteine (tHcy) concentration is associated with several multifactorial disorders including vascular disease, venous thrombosis (VT), neural tube defects (NTDs), Alzheimer's disease, schizophrenia, and cancer. The mechanisms underlying the association between plasma tHcy concentration and disease are unclear. Elucidation of the genetic aetiology of plasma tHcy-related diseases can aid in the understanding of disease aetiology and the development of diagnostic tools and therapeutic measures but is difficult due to the aetiologic complexity of multifactorial diseases. Genetic analysis of plasma tHcy concentration is easier and promotes the elucidation of the underlying pathology of homocysteine-related diseases. However, despite extensive research, which was mainly focussed on single locus effects of a limited number of candidate genes, the genetic aetiology of plasma tHcy is unresolved. Application of advanced genetic epidemiological study designs and analyses may promote the elucidation of the genetic background of plasma tHcy concentration and related diseases. The main objective of this thesis was to identify DNA variants that influence plasma tHcy concentrations and NTDs and VT by use of genetic epidemiological studies **(part 1)**. The second objective of this thesis was to develop and evaluate genetic epidemiological tools that facilitate the identification and characterization of genetic risk factors **(part 2)**.

Chapter 1 contains a general introduction on this thesis and includes a description of homocysteine metabolism, the relation between plasma tHcy concentration and disease, the role for plasma tHcy as intermediate phenotype in genetic studies for complex multifactorial diseases, current knowledge on the genetic aetiology of plasma tHcy and plasma tHcy-related genetic determinants of NTD and VT, and genetic epidemiological designs and analyses that may enhance the elucidation of genetic risk factors. Also, the objectives and outline of this thesis are given.

Part 1: Genetic determinants of homocysteine and related diseases

In **chapter 2** we aimed to map quantitative trait loci for plasma tHcy by performing a linkage analysis in 13 extended families comprising 264 individuals with measurement of plasma tHcy and 377 genetic polymorphic markers spread over the autosomal genome. The estimated heritability of age and sex adjusted plasma tHcy in this population was 44%. We identified one area on chromosome 16q12 with suggestive linkage evidence (LOD score 1.76). Two areas with weak linkage evidence were located on chromosome 12q14 (LOD score 1.57) and 13q31 (LOD score 1.52). A database search indicated that the region on 12q14 harboured speculative potential candidate genes: serine hydroxymethyltransferase 2 (*SHMT2*) and methyltransferase like 1 (*METTL1*).

In **chapter 3** we tried to identify genetic determinants of the one-carbon intermediate metabolites folate, homocysteine, and methionine using an extensive candidate-gene association strategy. We successfully genotyped 79 DNA variants in coding and non-coding regions in 40 one-carbon metabolism-related genes using a SNP-microarray approach in 190 unrelated Caucasian individuals. We classified the 40 genes in 5 metabolic processes: folate cycle, remethylation, transmethylation, transsulfuration, other. We applied single-locus association analysis and a multi-locus set-based analysis for the 5 gene subsets. SNP rs1801133 (*MTHFR*677C>T) was main determinant of serum folate (nominal *P*=0.002); other strong determinants included rs8101626 in *DNMT1* (nominal *P*=0.003) and rs1801394 (*MTRR*66A>G) (nominal *P*=0.011). *MTHFR*677C>T was also nominally associated to fasting and post-load plasma tHcy (nominal *P*=0.026 and 0.006, respectively). SNP rs2276598 (*DNMT3*A1523G>A) showed nominal association to fasting tHcy (nominal *P*=0.035). *CBS*844_845ins(68bp) was significant determinant of post-load tHcy (nominal P=0.0005; family-wise P=0.044) and explained 6.3% of its variance. A non-synonymous variant (rs672346) in *BHMT* (nominal *P*=0.014) and 3′ UTR rs1078004 in *ICMT* (nominal *P*=0.030) were related to methionine. Substantial overlap in strong DNA variants for fasting and post-load plasma tHcy was present. Only *MTHFR*677C>T showed strong association to more than one of the three metabolites. In conclusion, our extensive candidate-gene study resulted in establishment of known and identification of new candidate DNA variants for folate, homocysteine, and methionine concentrations.

Chapter 4 describes the analysis of 36 gene variants (of which most are known or suspected to affect function of the encoded protein) in 19 candidate genes related to homocysteine metabolism using haplotype and logic regression analysis in 461 unrelated Caucasian individuals. Our goal was to identify and characterize multi-locus genetic determinants of plasma tHcy concentration. For completeness, single locus association results were presented too. These were in line with our previous publications and showed a main role with modest marginal effects for *CBS*844_845ins(68bp), *CBS* 31bp VNTR, *COMT*324G>A, *MTHFR*677C>T and potential minor roles for other DNA variants in plasma tHcy. No haplotype effects in addition to the single locus effects were found. Fourteen DNA variants and 182 individuals were included in logic regression analysis. This analysis identified the following logical genotype combination: {*CBS* 31bp VNTR 18-18 and *FOLH1*1561CC and (*MTHFD*2011GG or *BHMT*595GA) and *MTHFR*677CT/TT}; it was present in 10% of the study population and associated with 5 µmol/L higher median fasting plasma tHcy. It explained ~17% of plasma tHcy variance in our population; the 5 DNA variants separately explained ~5% of variance.

In **chapter 5** we aimed at identifying genetic determinants related to folate and its metabolic pathways that are involved in the aetiology of spina bifida, the most common type of NTDs, using an extensive candidate-gene association study. We analysed 87 DNA variants in 45 genes in 180 patients and 190 controls based on application of

the SNP-array described in chapter 3. For those variants that showed nominal association to spina bifida, the relation with age and sex adjusted folate, cobalamin, and tHcy concentration was evaluated in the control sample. The strongest association with spina bifida that remained statistically significant after correction for multiple testing was found for rs1907362 in *CUBN*, encoding the intrinsic factor-cobalamin receptor (or cubilin). This intronic SNP was associated with decreased spina bifida risk and increased red blood cell (RBC) folate (P=0.001) and cobalamin levels (P=0.039). The second strongest association was found for the intronic rs2295809 in *TRDMT1*, encoding one of the many methyltransferases, which showed association to increased RBC folate too. Other top findings included variants in *ALDH1L1*, *SARDH*, and *SLC19A1*. No haplotype effects on top of main effects of single DNA variants were found. The identified multilocus logic regression model included rs1907362 in *CUBN* and rs2295809 in *TRDMT1* as separate variables and we did not find statistically significant interactions between the two variants.

In **chapter 6** we evaluated whether a 45 bp deletion/insertion variant in candidate gene *UCP2*, a potential regulator of mitochondrial reactive oxygen species production and oxidative stress, was associated with elevated plasma tHcy and increased risk for recurrent venous thrombosis (RVT) in 161 RVT cases and 386 controls. We did not find an association between fasting plasma tHcy and the variant. We did however find a positive association between the *UCP2* 45 bp insertion/deletion and post-load plasma tHcy in controls (P-value ANOVA 0.02) and indications for association in RVT cases (P-value ANOVA 0.09). The frequency of the *UCP2* 45 bp ins/ins genotype was 12.4% in RVT cases compared to 8.3% in controls (ins/ins vs. del/del: OR 1.8 (95% CI 1.0 to 3.4)).

Part 2: Genetic epidemiological designs and analyses

Chapter 7 presented a new hybrid design augmenting case-parent triads and control-mother dyads as population-based controls. Comparison to competing designs (i.e. case-parent triad, case-parent triad/control parents hybrid, and case-mother/control-mother dyad designs) showed that it is a powerful approach for testing both offspring and maternal genetic relative risk parameters under various scenarios, though it is slightly less powerful than the hybrid design that uses control parents. In addition, it allows testing of absence of population stratification bias, an underlying assumption in combining case and control data, and presence of mating symmetry. Even if the assumptions of absence of population stratification bias and presence of mating symmetry are not fulfilled, the hybrid design allowed for valid estimation of offspring and maternal genetic relative risk parameters.

In **chapter 8** we evaluated the application of traditional logistic regression analysis, the multifactor dimensionality reduction (MDR) method, logic regression analysis, and the set-association method for the identification of statistically interacting DNA variants. We simulated data for 200 cases and 200 controls for a null model of no genetic effects, five 2-locus statistical interaction models, and one model containing two 2-locus interactions, all defined using the penetrance scale, under four different minor allele frequencies. We found that the performance of logistic regression analysis was dependent on the underlying model and the application of multiple comparisons adjustment procedures. The ability to identify the causal loci was generally good for the other methods. All methods had difficulty identifying the two 2-locus interaction models with small or no marginal effects. Several practical and methodological characteristics of the methods that we encountered were reviewed.

In conclusion, extensive association studies for over 40 one-carbon metabolism-related genes identified novel candidates that warrant follow-up research and known DNA variants that had a small contribution in plasma tHcy variation each. Application of non-parametric multi-locus analysis allowed the identification of a strongly associated five-locus interacting genotype that explained 17% of plasma tHcy variation in our population. The candidate-gene studies also underlined the relative importance of *MTHFR*677C>T and *CBS* variants. The simultaneous analysis of plasma tHcy as well as folate and methionine concentrations indicated dissimilarities in genetic determinants for the three metabolites. The latter is important in view of elucidation of complex disease mechanisms, especially since the evidence that plasma tHcy itself may not be a causal element in disease pathology is accumulating. The simultaneous evaluation of genetic associations for plasma tHcy ánd for RVT or NTD led to identification of potential genetic risk factors for these diseases and also indicated whether or not the disease-mechanism may be plasma tHcy-related. The evaluation of four multi-locus analysis techniques for identification of interacting loci may be used to guide and facilitate the analysis and interpretation of multi-locus data as for our studies described in chapters 4 and 5 of this thesis. The new powerful and flexible hybrid design for identification of both maternal and offspring genotypic effects can be applied in future studies into congenital disorders like NTDs.

A discussion of the studies of this thesis and future research perspectives can be found in **chapter 9**.

Samenvatting

Homocysteïne is een zwavelhoudende intermediair in de folaatcyclus en het methionine metabolisme, ook wel bekend als de one-carbon metabolisme route. Een verhoogde concentratie van plasma totaal homocysteïne (tHcy) is geassocieerd met diverse multifactoriële aandoeningen, waaronder hart- en vaatziekten, veneuze trombose (VT), neurale buis defecten (NBD), de ziekte van Alzheimer, schizofrenie, en kanker. De onderliggende mechanismen zijn echter onduidelijk.

Opheldering van de genetische etiologie van de bovengenoemde aandoeningen geeft meer inzicht in de pathofysiologische processen en kan leiden tot de ontwikkeling van preventieve en therapeutische interventies maar is moeilijk vanwege de etiologische complexiteit van multifactoriële aandoeningen. Genetische analyse van plasma tHcy concentratie is makkelijker en bevordert de opheldering van de onderliggende pathologie van ziekten gerelateerd aan homocysteïne. Ondanks het feit dat er veel onderzoek naar genetische determinanten van plasma tHcy concentratie is verricht, met name naar effecten van een enkele variant in een beperkt aantal kandidaatgenen, is de genetische etiologie van plasma tHcy concentratie niet opgehelderd. Toepassing van geavanceerde genetisch epidemiologische studie designs and analyses kan deze opheldering van de genetische achtergrond van homocysteïne en gerelateerde aandoeningen bevorderen.

Het doel van het onderzoek beschreven in dit proefschrift was het identificeren van genetische varianten die plasma tHcy concentraties, en ook het risico op NBD en VT, beïnvloeden aan de hand van genetisch epidemiologische studies **(deel 1)**. Het tweede doel van dit proefschrift was het ontwikkelen en evalueren van genetisch epidemiologische designs en analyses die de identificatie en karakterisering van genetische determinanten bevorderen **(deel 2)**.

Hoofdstuk 1 omvat een algemene introductie en beschrijft het homocysteïne metabolisme, de relatie tussen plasma tHcy concentraties en ziekte, de rol voor plasma tHcy concentraties als intermediair fenotype in genetische studies naar multifactoriële aandoeningen, de huidige kennis omtrent genetische determinanten van plasma tHcy concentraties en aan homocysteïne gerelateerde determinanten van NBD en VT, en een beschrijving van genetisch epidemiologische studie designs en analyses die de opheldering van genetische determinanten bevorderen. Tevens worden de doelen en inhoud van dit proefschrift beschreven.

Deel 1: Genetische determinanten van homocysteïne en gerelateerde aandoeningen

In **hoofdstuk 2** beoogden we de lokalisatie van 'quantitative trait loci' voor plasma tHcy concentratie middels een koppelingsanalyse in 13 families bestaande uit 264

individuen waarin plasma tHcy concentraties en 377 polymorfe markers, verspreid over het autosomale genoom, zijn gemeten. De geschatte heritability van plasma tHcy concentratie, gecorrigeerd voor leeftijd en geslacht, was 44% in deze populatie. We vonden één gebied op chromosoom 16q12 met suggestief bewijs voor koppeling (LOD score 1.76). Twee gebieden met zwak bewijs voor koppeling waren gelokaliseerd op chromosoom 12q14 (LOD score 1.57) en 13q31 (LOD score 1.52). Databank onderzoek indiceerde potentiële kandidaatgenen in de 12q14 regio: serine hydroxymethyltransferase 2 (*SHMT2*) en methyltransferase like 1 (*METTL1*).

In **hoofdstuk 3** probeerden we genetische determinanten van folaat, homocysteïne, en methionine, intermediaire metabolieten van het one-carbon metabolisme, te identificeren aan de hand van een uitgebreide kandidaatgen associatie studie. We hebben 79 DNA varianten in coderende en niet-coderende regionen in 40 one-carbon metabolisme gerelateerde genen succesvol gemeten met behulp van een SNP microarray aanpak in 190 ongerelateerde Kaukasische individuen. We classificeerden de 40 genen in 5 metabole processen: folaatcyclus, remethylering, transmethylering, transsulfurering, overig. We hebben een single-locus associatie analyse en een multi-locus set-based analyse voor de 5 genklassen toegepast. SNP rs1801133 (*MTHFR*677C>T) was sterkste determinant van serum folaat (nominale *P*=0.002); overige sterke determinanten waren rs8101626 in *DNMT1* (nominale *P*=0.003) en rs1801394 (*MTRR*66A>G) (nominale *P*=0.011). *MTHFR*677C>T was ook nominaal geassocieerd met nuchter en post methionine-load plasma tHcy (nominale *P* respectievelijk 0.026 en 0.006). SNP rs2276598 (*DNMT3A*1523G>A) liet nominale associatie met nuchter plasma tHcy zien (nominale *P*=0.035). *CBS*844_845ins(68bp) was een significante determinant van post-load plasma tHcy (nominale *P*=0.0005; family-wise *P*=0.044) en verklaarde 6.3% van de variantie. Een non-synonieme variant (rs672346) in *BHMT* (nominale *P*=0.014) en 3' UTR rs1078004 in *ICMT* (nominale P=0.030) waren gerelateerd aan methionine. Substantiële overlap in sterke DNA varianten voor nuchter en post-load plasma tHcy was aanwezig. Alleen *MTHFR*677C>T liet sterke associatie zien met meer dan één van de drie metabolieten. Kortom, onze uitgebreide kandidaatgen studie resulteerde in constatering van reeds bekende en de identificatie van nieuwe kandidaat DNA varianten voor folaat, homocysteïne, en methionine concentraties.

Hoofdstuk 4 beschrijft de analyse van 36 gen varianten (waarvan voor de meeste bekend is of wordt vermoed dat ze de functie van het gecodeerde eiwit beïnvloeden) in 19 kandidaatgenen gerelateerd aan het homocysteïne metabolisme met behulp van haplotype en logic regressie analyse in 461 ongerelateerde Kaukasische individuen. Ons doel was het identificeren en karakteriseren van multi-locus genetische determinanten van plasma tHcy concentratie. Voor de volledigheid werden ook resultaten van single-locus associaties gepresenteerd. Deze kwamen overeen met onze eerder gepubliceerde resultaten en lieten een hoofdrol zien, met bescheiden marginale effecten, voor *CBS*844_845ins(68bp), *CBS* 31bp VNTR, *COMT*324G>A, *MTHFR*677C>T en

potentiële kleine rollen voor andere DNA varianten voor plasma tHcy. Er werden geen haplotype effecten gevonden onafhankelijk van de single- locus effecten. Veertien DNA varianten en 182 individuen waren geïncludeerd in de logic regressie analyse. Deze analyse identificeerde de volgende Booleaanse of logische genotype combinatie: {*CBS* 31bp VNTR 18-18 en *FOLH1*1561CC en (*MTHFD*2011GG of *BHMT*595GA) en MTHFR677CT/TT}; deze combinatie was aanwezig in 10% van de studiepopulatie en geassocieerd met een 5 µmol/L hogere mediane nuchtere plasma tHcy concentratie. De multi-locus combinatie verklaarde ~17% van de variantie in plasma tHcy in onze populatie; de 5 DNA varianten afzonderlijk verklaarde ~5% variantie.

In **hoofdstuk 5** richtten we ons op het identificeren van genetische determinanten van folaat en de metabole pathways die zijn betrokken in de etiologie van spina bifida, het meest voorkomende type NBD, middels een uitgebreide kandidaatgen associatie studie. We analyseerden 87 DNA varianten in 45 genen in 180 patiënten en 190 controles met gebruikmaking van de SNP microarray die is beschreven in hoofdstuk 3. Voor die varianten die nominale associatie lieten zien met spina bifida, werd de associatie met voor leeftijd- en geslacht- gecorrigeerde concentraties van folaat, cobalamine, en tHcy geëvalueerd in de controles. De sterkste associatie met spina bifida, die statistisch significant bleef na correctie voor de meervoudige toetsing, werd gevonden voor rs1907362 in *CUBN*, coderend voor de intrinsic factor-cobalamin receptor (of cubiline). Deze SNP, gelegen in een intron, was geassocieerd met een verlaagd risico voor spina bifida en verhoogde concentraties folaat in rode bloed cellen (RBC) (P=0.001) en cobalamine (P=0.039). De één na sterkste associatie werd gevonden voor rs2295809 gelegen in een intron in *TRDMT1*, coderend voor één van de vele methyltransferases, welke ook een associatie met verhoogd RBC folaat liet zien. Andere top bevindingen betroffen varianten in *ALDH1L1*, *SARDH*, en *SLC19A1*. Er werden geen haplotype effecten gevonden die onafhankelijk waren van marginale effecten van single-locus varianten. Het geïdentificeerde multi-locus logic regressie model bevatte rs1907362 in *CUBN* en rs2295809 in *TRDMT1* als aparte variabelen en we vonden geen statistisch significante interactie tussen deze twee varianten.

In **hoofdstuk 6** evalueerden we of de 45 bp deletie/insertie variant in het kandidaatgen *UCP2*, een potentiële regulator van productie van mitochondriële reactieve zuurstof radicalen en oxidatieve stress, was geassocieerd met een verhoogd plasma tHcy en verhoogd risico voor een recidief veneuze trombose (RVT) in 161 RVT patiënten en 386 controlepersonen. We vonden geen associatie tussen nuchter plasma tHcy en de variant. We vonden echter wel een positieve associatie voor de *UCP2* 45 bp insertie/deletie and post-load plasma tHcy in controles (P ANOVA 0.02) en aanwijzingen voor een associatie in RVT patiënten (P ANOVA 0.09). The frequentie van het *UCP2* 45 bp ins/ins genotype was 12.4% in RVT patiënten vergeleken met 8.3% in controles (ins/ins vs. del/del: OR 1.8 (95% CI 1.0 tot 3.4)).

Deel 2: Genetisch epidemiologische designs en analyses

In **hoofdstuk 7** presenteerden we een nieuw hybride design dat patiënt-ouder trios uitbreidt met controle-moeder duos als populatie controles. Een vergelijking met concurrerende designs (d.w.z. patiënt-ouder trio design, patiënt-ouder trio/controle ouders hybride design, en patiënt-moeder/controle-moeder duo design) liet zien dat het een krachtige aanpak is voor het testen van de genetische relatieve risico parameters voor zowel kind als moeder in verschillende scenario's, alhoewel het iets minder power heeft dan het hybride design dat ouders van controles gebruikt. Tevens is met het nieuwe hybride design mogelijk om de afwezigheid van populatie stratificatie bias te testen, een onderliggende assumptie in het combineren van patiënt en controle data, en de aanwezigheid van 'mating symmetrie'. Zelfs als niet wordt voldaan aan de assumptie van afwezigheid van populatie stratificatie bias en aanwezigheid van mating symmetrie, laat het hybride design een valide schatting van kind- en maternale genetische relatieve risico parameters toe.

In **hoofdstuk 8** evalueerden we de toepassing van traditionele logistische regressie analyse, de 'multifactor dimensionality reduction' (MDR) methode, logic regressie analyse, en de set-associatie methode voor de identificatie van DNA varianten die statistische interactie vertonen. We simuleerden data voor 200 patiënten en 200 controles voor een nul-model (geen genetische effecten), vijf 2-locus statistische interactie modellen, en één model dat twee 2-locus interacties bevatte, allen gedefinieerd op basis van penetrantie, voor vier verschillende minor allel frequenties (MAF). We vonden dat de prestatie van logistische regressie analyse afhankelijk was van het onderliggende model en de toepassing van procedures voor de correctie voor meervoudige toetsing. Het vermogen om de causale loci te identificeren was over het algemeen goed voor de andere methoden. Alle methoden hadden moeite met het identificeren van de twee 2-locus interactie modellen met kleine of geen marginale effecten. Diverse praktische en methodologische karakteristieken die werden ervaren in de toepassingen van de verschillende methoden, werden besproken.

Samenvattend kunnen we zeggen dat de uitgebreide associatie studies voor meer dan 40 one-carbon metabolisme gerelateerde genen heeft geleid tot identificatie van nieuwe kandidaten voor vervolgonderzoek en het vinden van reeds bekende DNA varianten die elk een klein deel van de variatie in plasma tHcy verklaarden. Toepassing van non-parametrische multi-locus analyse maakte de identificatie van een sterk geassocieerd vijf-locus genotype mogelijk dat 17% van de plasma tHcy variatie in onze populatie verklaarde. De kandidaatgen studies benadrukten ook het relatief grote belang van *MTHFR*677C>T en *CBS* varianten. De simultane analyse van plasma tHcy en folaat en methionine concentraties liet verschillen in genetische determinanten zien voor deze drie metabolieten. Dit laatste is belangrijk in het kader van de opheldering

van pathofysiologische mechanismen voor multifactoriële ziekten, met name gezien het toenemende bewijs dat plasma tHcy niet zelf een causale factor is. De simultane evaluatie van genetische associaties voor plasma tHcy en voor RVT of NTD leidde tot de identificatie van potentiële genetische risicofactoren voor deze aandoeningen en gaven ook aanwijzingen over het al dan niet gerelateerd zijn aan plasma tHcy van het onderliggende ziektemechanisme. De evaluatie van vier multi-locus analysetechnieken ter identificatie van interacterende loci kan gebruikt worden ter begeleiding en ondersteuning van de analyse en interpretatie van multi-locus data zoals ook gedaan is in de studies beschreven in hoofdstukken 4 and 5 van dit proefschrift. Het nieuwe krachtige en flexibele hybride design voor de identificatie van genotype effecten van zowel moeder als kind kan in de toekomst worden toegepast in studies naar aangeboren afwijkingen zoals NBD.

Een discussie van de studies beschreven in dit proefschrift en mogelijkheden voor vervolgonderzoek worden gegeven in **hoofdstuk 9**.

Dankwoord

Dit proefschrift is tot stand gekomen middels een interactie tussen veel mensen met diverse achtergronden. Een geslaagde synergie wat mij betreft! Ik wil de betrokkenen graag hier bedanken.

Dr. den Heijer, beste Martin, dit proefschrift is mogelijk gemaakt door jouw VENI subsidie. Ik wil je graag bedanken voor je begeleiding en de ruimte en mogelijkheden die je me in de afgelopen jaren hebt gegeven om mezelf nieuwe kennis en onderzoeksvaardigheden eigen te maken en me in te zetten op meerdere afdelingen binnen het UMC St Radboud. Ik heb bewondering voor en geleerd van je enthousiaste en creatieve kijk op het (homocysteïne)onderzoek en je vermogen om naast al het klinische werk nog bergen succesvol werk op het gebied van onderzoek te verzetten.

Dr. Blom, beste Henk, de regelmaat waarmee wij elkaar zagen, heeft enorm gevarieerd over de afgelopen jaren. Je hebt me de nodige kennis omtrent homocysteïne bijgebracht en veel van de metingen die beschreven staan in dit proefschrift zijn verricht door mensen van jouw lab. Ik heb me altijd vrij gevoeld om bij je aan te kloppen en weer eens een vraag bij je neer te leggen, of het nu ging om een homocysteïnegerelateerde zaak of een andere kwestie. Bedankt voor je open deur.

Prof. Kiemeney, beste Bart, vanaf het begin van mijn aanstelling heb je op de achtergrond meegedacht en met name ook vooruitgedacht over de tijd na mijn promotie. Ik wil je bedanken voor je bijdrage aan dit proefschrift en je steun die me ook heeft gebracht op de plek waar ik nu zit. Het laatste jaar zijn we intensiever samen gaan werken maar het echte werk moet nog beginnen. Ik kijk er naar uit.

Prof. Hermus, beste Ad, ik ben in 2004 begonnen als junior onderzoeker op jouw afdeling. Bedankt dat ik in de afgelopen jaren heb kunnen rekenen op je interesse en snelle correcties van manuscripten, ook na mijn verhuizing naar de afdeling Epidemiologie.

During my PhD training, I spent three months at the Statistical Genetics Group, SGDP, IOP, King's College, London, UK. Under supervision of dr. Knight and prof. Sham I worked on the study described in chapter 8 of this PhD thesis. Dr. Knight, dear Jo, I've truly enjoyed our meetings, talks and laughter. Thank you very much! Dear prof. Sham, thank you for your inspiring supervision during the two weeks I spent at your department at the University of Hong Kong. I would also like to thank Frühling Rijsdijk, Desmond Campbell, Ben Neale, Mei-Hua Hall, and Mark Knight for scientific as well as social support. I look back at my three months at the Statistical Genetics Group with great pleasure and hope to be able to visit you again in the future.

In 2007 I spent one month at the Biostatistics Branch, NIEHS, NIH, Research Triangle Park, North Carolina, USA. Dear dr. Weinberg, dear Clare, dear dr. Umbach, dear David, dear dr. Shi, dear Min, your stimulating and efficient supervision has led to the manuscript presented in chapter 7 of my PhD thesis. Thank you very much for your warm

welcome and all the time you've spent on discussing the designs, LEM, and 'degrees of freedom'. I look back on a very pleasant, fruitful collaboration and instructive time.

Homocysteine Research Group, jullie werk, gedrevenheid en kennis hebben bijgedragen aan het resultaat dat hier ligt. Specifieke dank gaat uit naar Sandra Heil, Henkjan Gellekink, en Ivon van der Linden. Ik wil hier ook Gerly van der Vleuten en Jacqueline de Graaf noemen. Ik ben blij dat we onze gedeelde onderzoeksinteresses hebben kunnen omzetten in een aantal gezamenlijke publicaties.

Beste (oud)collega's van de afdeling Endocriene ziekten en afdeling Epidemiologie, Biostatistiek en HTA, met mijn onderzoek naar de genetische epidemiologie van homocysteïne was ik een vreemde eend in de bijt op beide afdelingen. Ik heb echter met plezier gewerkt bij Endo en voel me op mijn plek bij EBH. Dank voor jullie collegialiteit en de interesse die jullie hebben getoond in mijn promotieonderzoek!

Beste collega's en andere betrokkenen van het lab Multifactorieel, afdeling Antropogenetica, het bijwonen van de werkbesprekingen en het mee mogen werken aan een groot aantal projecten hebben meer bijgedragen aan dit proefschrift en aan mijn ontwikkeling als genetisch epidemioloog dan jullie misschien wel denken! Dank! Dr. Franke, beste Barbara, bedankt voor je vertrouwen in mij tijdens mijn eerste stappen als genetisch epidemioloog. Ik ben blij dat ik je beter heb mogen leren kennen tijdens onze gezamenlijke tijd in Londen. Dr. Coenen, beste Marieke, onze samenwerking is in het laatste jaar hechter geworden. Ik kijk uit naar een verdere uitdieping van onze gemeenschappelijke en aanvullende onderzoeksinteresses.

Lotte Knapen en Ivon van der Linden, lieve paranimfen! Onze eerste ontmoeting viel praktisch samen met de dag dat ik aan dit promotie-avontuur begon. Jullie zijn in de afgelopen jaren erg belangrijk voor me geworden. Ik vind het dan ook enorm fijn en bijzonder dat jullie op de dag van mijn promotie naast me willen staan.

Beste vrienden, jullie mogen me eindelijk dr. Sita gaan noemen! Bedankt voor al die keren dat jullie mijn diepte- en hoogtepunten en alles daar tussen in met me hebben gedeeld; mijn speciale dank gaat hierbij uit naar Koen. De met jullie ondernomen activiteiten (of dit nu gepaard ging met feestjes en kroegbezoek, thee en koekjes, reizen, sportieve of culturele aangelegenheden, of anders) en onze gesprekken hebben me enorm gemotiveerd in de afgelopen promotiejaren. Ik kijk uit naar onze toekomstige 'samenwerking'!

Papa, mama, en ook Danny en Chantal, dank voor jullie onvoorwaardelijke steun. We hebben het vaak gehad over de vorderingen met betrekking tot het proefschrift en ik ben trots en blij dat ik jullie eindelijk kan zeggen: het is af!

About the author

Sita Hendrika Henriette Maria Vermeulen was born on the 15th of June 1978 in Appeltern, the Netherlands. From 1990 till 1996 she attended secondary school at the Titus Brandsma Lyceum in Oss. In 1998 she started with the study Nutrition and Health at Wageningen University. During her university training she performed an internship at the Department of Hygiene and Public Health at the University of Cagliari, Italy, and at the Department of General Practice, Academic Medical Center, Amsterdam. After obtaining her master degree with honours in November 2002, she started working at the latter department on a research project on the generalizability of phase III clinical trials. From February 2004 until February 2009 she worked on her PhD project at the Department of Endocrinology of the Radboud University Nijmegen Medical Centre (RUNMC) of which the results are described in this thesis. During this time, Sita was also appointed at the Departments of Human Genetics (dr. B. Franke) and Epidemiology, Biostatistics and HTA (prof. L. Kiemeney) at the RUNMC as genetic epidemiologist and teacher. In the second year of her PhD project Sita received a 'Frye Stipend' for talented female researchers from the Radboud University Nijmegen. She visited the Statistical Genetics Group at the SGDP, King's College in London, UK (dr. J. Knight, prof. P. Sham) and the Biostatistics Branch at the NIEHS, NIH in Research Triangle Park, North Carolina, USA (dr. C. Weinberg). Sita has continued her work at the Department of Epidemiology, Biostatistics, and HTA and at the Department of Human Genetics and has focussed on the genetic epidemiology of cancer and pharmacogenetics.

List of Publications

Radstake TR, Sweep FC, Welsing P, Franke B, **Vermeulen SH**, Geurts-Moespot A, Calandra T, Donn R, van Riel PL. Correlation of rheumatoid arthritis severity with the genetic functional variants and circulating levels of macrophage migration inhibitory factor. Arthritis Rheum. 2005;52:3020-3029.

van Beynum IM, Kapusta L, den Heijer M, **Vermeulen SH**, Kouwenberg M, Daniels O, Blom HJ. Maternal MTHFR 677C>T is a risk factor for congenital heart defects: effect modification by periconceptional folate supplementation. Eur Heart J. 2006;27:981-987.

Vermeulen SH, van der Vleuten G, de Graaf J, Hermus AR, Blom HJ, Stalenhoef AF, den Heijer M. A genome-wide linkage scan for homocysteine levels suggests three regions of interest. J Thromb Haemost. 2006;4:1303-1307.

van der Linden IJ, den Heijer M, Afman LA, Gellekink H, **Vermeulen SH**, Kluijtmans LA, Blom HJ. The methionine synthase reductase 66A>G polymorphism is a maternal risk factor for spina bifida. J Mol Med. 2006;84:1047-1054.

Vermeulen SH, den Heijer M, Sham P, Knight J. Application of multi-locus analytical methods to identify interacting loci in case-control studies. Ann Hum Genet. 2007;71:689-700.

Gellekink H, Muntjewerff JW, **Vermeulen SH**, Hermus AR, Blom HJ, den Heijer M. Catechol-O-methyltransferase genotype is associated with plasma total homocysteine levels and may increase recurrent venous thrombosis risk. Thromb Heam. 2007;98:1226-1231.

Kiemeney L, Aben K, **Vermeulen S.** Nieuwe technologie onderzoek erfelijke risicofactoren. Tijdschrift Kanker. December 2007.

Anney RJ, Hawi Z, Sheehan K, Mulligan A, Pinto C, Brookes KJ, Xu X, Zhou K, Franke B, Buitelaar J, **Vermeulen SH**, Banaschewski T, Sonuga-Barke E, Ebstein R, Manor I, Miranda A, Mulas F, Oades RD, Roeyers H, Rommelse N, Rothenberger A, Sergeant J, Steinhausen HC, Taylor E, Thompson M, Asherson P, Faraone SV, Gill M. Parent of origin effects in attention/deficit hyperactivity disorder (ADHD): Analysis of data from the international multicenter ADHD genetics (IMAGE) program. Am J Med Genet B Neuropsychiatr Genet. 2008;147B:1495-1500.

Fliers E, Rommelse N, **Vermeulen SH**, Altink M, Buschgens CJ, Faraone SV, Sergeant JA, Franke B, Buitelaar JK. Motor Coordination Problems in Children and Adolescents with ADHD: effects of Age and Gender. J Neural Transm. 2008;115:211-220.

Kooij JS, Boonstra AM, **Vermeulen SH**, Heister AG, Burger H, Buitelaar JK, Franke B. Response to methylphenidate in adults with ADHD is associated with a polymorphism in SLC6A3 (DAT1). Am J Med Gen B Neuropsychiatr Genet. 2008;147:201-208.

Verhaegh GW, Verkleij L, **Vermeulen SH**, den Heijer M, Witjes JA, Kiemeney LA. Polymorphisms in the H19 Gene and the Risk of Bladder Cancer. Eur Urol. 2008;54:1118-1126.

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, Pálsson S, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, **Vermeulen SH**, Goldstein AM, Tucker MA, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Two newly identified genetic determinants of pigmentation in Europeans. Nat Genet. 2008;40:835-837.

Heil SG, **Vermeulen SH**, Van der Rijt-Pisa BJM, den Heijer M, Blom HJ. Role for mitochondrial uncoupling protein-2 (UCP2) in hyperhomocysteinemia and venous thrombosis risk? Clin Chem Lab Med. 2008;46:655-659.

Kiemeney LA, Thorlacius S, Sulem P, Geller F, Aben KK, Stacey SN, Gudmundsson J, Jakobsdottir M, Bergthorsson JT, Sigurdsson A, Blondal T, Witjes JA, **Vermeulen SH**, Hulsbergen-van de Kaa CA, Swinkels DW, Ploeg M, Cornel EB, Vergunst H, Thorgeirsson TE, Gudbjartsson D, Gudjonsson SA, Thorleifsson G, Kristinsson KT, Mouy M, Snorradottir S, Placidi D, Campagna M, Arici C, Koppova K, Gurzau E, Rudnai P, Kellen E, Polidoro S, Guarrera S, Sacerdote C, Sanchez M, Saez B, Valdivia G, Ryk C, de Verdier P, Lindblom A, Golka K, Bishop DT, Knowles MA, Nikulasson S, Petursdottir V, Jonsson E, Geirsson G, Kristjansson B, Mayordomo JI, Steineck G, Porru S, Buntinx F, Zeegers MP, Fletcher T, Kumar R, Matullo G, Vineis P, Kiltie AE, Gulcher JR, Thorsteinsdottir U, Kong A, Rafnar T, Stefansson K. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. Nat Genet. 2008;40:1307-1312.

Shi M, Umbach DM, **Vermeulen SH**, Weinberg CR. Making the most of case-mother/control-mother studies. Am J Epidemiol. 2008;168:541-547.

Vermeulen SH, Shi M, Weinberg CR, Umbach DM. A hybrid design: case-parent triads supplemented by control-mother dyads. Genet Epidemiol. 2009;33:136-144.

Franke B, **Vermeulen SH**, Coenen MJ, Steegers-Theunissen RPM, Schijvenaars MMVAP, Scheffer H, den Heijer M, Blom HJ. Analysis of 45 folate related genes in spina bifida: involvement of Cubilin (CUBN) and tRNA aspartic acid methyltransferase 1 (TRDMT1). Birth Defects Res A Clin Mol Teratol. 2009;85:216-226.

Fliers E, **Vermeulen S**, Rijsdijk F, Altink M, Buschgens C, Rommelse N, Faraone S, Sergeant J, Buitelaar J, Franke B. ADHD and Poor Motor Performance From a Family Genetic Perspective. J Am Acad Child Adolesc Psychiatry. 2009;48:25-34.

Kiemeney LA, Grotenhuis AJ, **Vermeulen SH**, Wu X. Genome-wide association studies in bladder cancer: first results and potential relevance. Curr Opin Urol. In Press.